Harald (H.) Zimmermann

Das Lexikon in der maschinellen Sprachanalyse

Druck: Offsetdruckerei Wolf, Heppenheim
Papier: Gebr. Buhl Papierfabriken KG, Ettlingen/Baden
Buchbinderei: Kränkl, Heppenheim
Printed in Germany
ISBN-3-7610-5707-5

VORWORT

Kaum irgendwo anders in der Sprachwissenschaft ist Teamarbeit so notwendig wie in dem Bereich der Computerlinguistik, der die Bewältigung großer Datenmengen (Regeln, Texte, Lexika) zum Ziel hat. Ohne die Vorarbeiten der Mitarbeiter der Saarbrücker Arbeitsgruppe für linguistische Datenverarbeitung, die im Jahre 1969 zur Erstellung eines maschinellen Syntax-Analysators (Parsers) führten, hätte diese Arbeit in der vorliegenden Form nicht entstehen können. Ihnen allen gilt mein Dank, besonders aber dem Leiter dieser Forschungsgruppe, Herrn Prof. Eggers, der die Arbeit betreut und gefördert hat.
H. H. Z.

INHALT

< S. II>

<S. III>

LITERATURVERZEICHNIS

Agricola, E. (Polysyntaktizität) IV/1
Syntaktische Mehrdeutigkeit (Polysyntaktizität) bei der Analyse des Deutschen und des Englischen. Berlin 1968.
Antal, L. (Dictionary)
A new type of dictionary. In : Linguistics 1 (1963) S. 75-84.
Bahr, J. (Lexikographie)
Technische Verfahren in der Lexikographie. In: Zs.f.dt. Sprache 22 (1966) S. 96 – 111.
Bar-Hillel, Y. (Translation)
The Present Status of automatic Translation of Languages. In: Advances in Computers 1 (1960) S. 91-163.
Berner, K.E. (Sprachmittlerwesen)
Das Sprachmittlerwesen. Der Sprachendienst der Bundeswehr (III). In: Wehrkunde 12 (1965) S. 650-653.
Bierwisch, M. (Strukturalismus)
Strukturalismus. Geschichte, Probleme und Methoden. In: Kursbuch 5 (1966) S. 77-152.
Billmeier, G. (Simulation)
Simulation verbalen Verhaltens. In: Vorabdruck der Vorträge der G.I.-Fachtagung: Information Retrieval Systeme, Management Information Systeme, Stuttgart, 9.-11.Dez. 1970, S. 102-106.
Borko, H. (Language)
(ed.) Automated Language Processing. New York 1967.
Borkowski, C. (Personal Names)
An Experimental System for Automatic Indentifieation of Personal Names and Personal Titles in Newspaper Texts. In: American Documentation 18 (1967) Nr. 3.
Botha, R.P. (Lexicon)
The Function of the Lexicon in Transformational Generative Grammar. Den Haag 1968.
Bünting, K.-D. (Morphologische Strukturen)
Morphologische Strukturen deutscher Wörter: Ein Problem der linguistischen Datenverarbeitung. (Dias.) Bonn 1969.
Chomsky, N. (Aspekte)
Aspekte der Syntax-Theorie. Frankfurt 1969. (Originaltitel: Aspects of the Theory of Syntax, Mir, 1965).
Chomsky, N. (Nominalization)
Remarks on Nominalization. Publikation des Indiana University Linguistics Club, 1968. Demnächst in: Jacobs, Rosenbaum (eds.), Readings in Tranabrmational Grammar.
Clyne, M. (Komposita)
Ökonomie, Mehrdeutigkeit und Vagheit bei Komposita in der deutschen Gegenwartssprache, insbesondere in der Zeitungssprache. In: Muttersprache 78 (1968) S. 122-126.
Dietrich, R. (Starke Verben)
Eine formale Beschreibung der Starken und Unregelmäßigen Verben der deutschen Gegenwartsprache. Linguistische Arbeiten des Germanistischen Instituts und des Institute für angewandte Mathematik der Universität des Saarlandes, Arbeitsbericht Nr. 9, Saarbrücken 1970 (Juni 1970).
Dietrich, R. (Studien)

Studien zur maschinellen Lemmatisierung verbaler Wortformen in Texten der Deutschen Gegenwartssprache. (Dias.) Saarbrücken 1971.

DUDEN (Grammatik)
Der Große DUDEN (Bd. 4) Grammatik der deutschen Gegenwartsprache. Mannheim 1966.

DUDEN (Rechtschreibung)
Der Große DUDEN (Bd. 1) Rechtschreibung der deutschen Sprache und der Fremdwörter. Mannheim 1961.

Dworatschek, S. (Datenverarbeitung)
Einführung in die Datenverarbeitung. Berlin

Eggers, H. (Syntaxanalyse)
(et. al.) Elektronische Syntaxanalyse der deutschen Gegenwartssprache. Ein Bericht. Tübingen 1969. (Mitarbeiter dieses Buches sind: R. Dietrich, W. Klein,

R. Rath, A. Rothkegel, H.J. Weber und H. Zimmermann Eggers, H. (Sprache)
Zur Syntax der deutschen Sprache der Gegenwart. In: Studium Generale 15 (1962) S. 49-59.

Engels, L.K. (Analyse)
Automatische Analyse van het Engels. In: Linguistica Antverpiensia 2 (1968) S. 177-187

Fillmore, Ch.J. (Case)
The Case for Case. In: Bach/Harms (eds.), Universale in Linguistin Theory. New York, S. 1-88.

Friedmann, J. (Application)
Application of a Computer System for Transformational Grammar. Preprint Nr. 14 der International Conference an Computational Linguistics in Sanga Säby bei Stockholm. (1969).

Garvin, P.L. (Computer)
Computer Participation ir Linguistic Research. In: Language 38 (1962) S. 385-389.

Gausl, I. (Vocabulary)
Vocabulary: Its Measurements and Growth. In Archives of Psychology 33 (1939) S. 5-52.

Gougenheim, G. (Ddtermination)
Sur la détermination du sens d'un mot au moyen du contexte. In: Traduction Automatique 2 (1961) S. 16-17.

Gougenheim, G. (Dictionnaire)
Dictionnaire fondamental de la langue française. Paris 1958.

Gougenheim, G. (Elaboration)
(et al.) L'Elaboration du frangais 4lgmentaire. Paris 1956.

Graff, E.G. (Sprachschatz)
Althochdeutscher Sprachschatz oder Wörterbuch der althochdeutschen Sprache. Berlin 1834-42.

Grimm, J. (Wörterbuch)
und W. Grimm: Deutsches Wörterbuch. Leipzig 1854-1954.

Gross, M. (Verbe)
Grammaiie transformationelle. Syntaxe du verbe. Paris 1968.

Gruber, J.S. (Relations)
Studies in Lexical Relations. Publikation des Indiana University Linguistics Club, 1970.

Hays, D.G. (Computational Linguistics)
Introduction to Cumputational Linguistics. New York 1967.

Heinrich, W. (Wörterbuch)

Die Herstellung eines syntaktischen Wörterbuchs der deutschen Sprache auf Magnetband. In: Beiträge zur Linguistik und Informationsverarbeitung 13 (1968) S. 43-57.

v.Held, W. (Phrasenstrukturgrammatik)
Eine verallgemeinerte Phrasenstrukturgrammatik mit einem Satzanalyse-Algorithmus. IPK-Forschungsbericht Bd. 33, Hamburg 1971.

Janda, J.W. (Dokumentation)
A. Rothkegel und H. Zimmermann, Dokumentation eines Programms zur Analyse russischer Sätze. In: Beiträge zur Linguistik und Informationsverarbeitung 19/20 (1970).

Jankowsky, K.R. (Lexicology)
On Scope and Methode of Lexicology. In: Orbis 18 (1969) S. 173-185.

Josselson, H.H. (Lexicography)
Lexicography and the Computer. In: In Honor Roman Jacobson, Bd. 3, Den Haag 1967, S. lo46-1059.

Kastowgky, D. (Wortbildung)
Wortbildung und Nullmorphem. In: Linguistische Berichte 2 (1969) S. 1-13.

King,R.D. (Grammar)
Historical linguistics and generative grammar. Englewood Cliffs 1969.

Klein, S. (Coding)
und R.F. Simmons: A Computational Approach to Grammatical Coding of English Words. In Journal of the ACM 10 (1963) S. 334-347.

Klein, W. (Analysegrammatik)
Eine Analysegrammatik. In: Linguistische Berichte 10 (1970) S.. 13-26.

Klein, W. (Studien)
Studien zur Theorie der Maschinellen Syntaktischen Analyse. (Dies.) Saarbrücken 1970.

Klein, W. (Trakl)
und H.H. Zimmermann: Lemmatisierter Index zu Georg Trakls Dichtungen. Indices zur deutschen Literatur 6, Frankfurt 1971.

Krallmann, D. (Datenbank)
Linguistische Datenbank und kumDlatives Wörterbuch. Bericht des Kolloquiums über Linguistische Datenverarbeitung (Mannheim, 26.- 27.Nov. 1968), Mannheim 1969.

Lamb, S.M. (Digital Computer)
The digital computer as an aid in linguistics. In: Language 37 (1961) S. 382-412.

LEMMATISIERUNG
Automatische Lemmatisierung. Ein Bericht von W. Klein und R. Rath unter Mitarbeit von M. Bartoli, R. Dietrich, A. Rothkegel, H.J. Weber und H. Zimmermann (= Linguistische Arbeiten des Germanistischen Instituts und des Instituts für Angewandte Mathematik der Universität des Saarlandes, Nr. 10),Juni 1971.

Lenders, W. (Lexical Systems)
Static and Dynamie Lexical Systems. In: IPK-Forschungsbericht 69/5, Hamburg 1969, S. 1-8.

Mater, E. (Wörterbuch)
Rückläufiges Wörterbuch der deutschen Gegenwartssprache. Leipzig 1965.

Meier, H, (Sprachstatistik)
Deutsche Sprachstatistik. Hildesheim 1964.

Neuhaus, J. (Wortableitungen)

Semantische und Phonologische Beschränkungen in der Grammatik der Wortableitungen. In: Stechow, A. von (ed.), Beiträge zur generativen Grammatik (=Schriften zur Linguistik 3), Braunschweig 1971, S. 178-183.

Pendergraft, E.D. (Languages)
Translating Languages. In: Borko (Language) S. 291-323.

Perschke, S. (Sprachübersetzung)
Die Anwendung der maschinellen Sprachübersetzung in der Dokumentation bei EURATOK. Nachrichten für Dokumentation 20 (1969) Nr. 5.

Pfeffer, J.A. (Grunddeutsch)
Grunddeutsch. Basic (Spoken) German Word List. Englewood Cliffs 1964.

Raskin, J.F. (Humanities)
Programming Languages for the H nities. In: Computers and the Humanities 5 (1971) S. 15-158.

Rath, R. (Lemmatisierung)
Vorschläge zur Automatischen Lemmatisierung (AL) deutscher Adjektive. In: Linguistische Berichte 12 (1971) S. 53-59.

Reifler, E. (Compounds)
Mechanical. Determination of the Cbnstituents of German Substantive Compounds. In: Machine Translation 2 (1955) S. 3-14.

Rothkegel, A.E. (Syntagmen)
Feste Syntagmen – Grundlagen, Strukturbeschreibung und automatische Analyse. (Dies.) Saarbrücken 1971 (in Vorbereitung).

Sammet, H.E. (Programming languages)
Programming Languages: History and Fundamentale. Englewood Cliffs 1969.

Schirmer, B. (Übersetzungssystem)
Erfahrungen bei der Entwicklung eines maschinellen Übersetzungssysteme (Englisch-Deutsch). In:IBM-Nachrichten 19 (1969) S. 600-605.

Schmitz, W. (Präpositionen)
Der Gebrauch der deutschen Präpositionen. München51968.

Schnelle, H. (Theorie)
Neue Aspekte in der Theorie des Übersetzens. In: Sprache im technischen Zeitalter 23 (1967) S- 239-248.

Schoene, M. (Mots)
Vie et mort des mots. Paris 1951.

Schweisthal, K.G. (Präpositionen)
Präpositionen in der maschinellen Sprachbearbeitung. (= Schriftenreihe zur Kommunikativen Grammatik, ed.A. Hoppe, Bd. 1) Bonn 1971.

Smith, F. (Genesis)
und G.A. Miller (ede.), The Genesis of Language. A Psycholinguistic Approach. Proceedings of a Conference on 'Language Development in Children'...,Cambridge (Mass.) 1966.

SPRACHE UND MASCHINEN
Computer in der Übersetzung und in der Linguistik. In: Sprache im technischen Zeitalter 23 (1967) S. 218-238.

Stolz, W.S. (Coding)
et. al.: A Stochastic Approach to the Grammatical Coding of English. In: Communications of the ACM 8 (1965) S. 399-405.

Tesnière, L. (Esquisse)
Esquisse d'une syntaxe strucurale. Paris 1953.

Tollenaere, F. de (Lexikographie)
Lexikographie mit Hilfe des elektrischen Informationswandlers. In: Zeitschrift für deutsche Sprache 21 (965) S. 1-19.

Toma, P. (SERNA-System)

SERNA-System. Washington, Georgetown University, 1959. Wahrig, G. (Wörterbuch)
Deutsches Wörterbuch, Gütersloh 1968.
Weber, H.J. (Homographie)
Bestimmung der Wortklassen. In: Eggers (Syntaxanalyse) S. 62-89.
Wissyn
Introduction to the Program. (Diese Programmbeschreibung (masch.) kann vom Mass. Comm. Research Center, University of Wisconsin, Madison, Wis. 53711 bezogen werden.)
Wenzel, G.(Textverarbeitung)
Textverarbeitung auf der Graphemebene. Diss. Stuttgart 1967.
Wunderlich, D. (Nominalisierungen)
Warum die Darstellung von Nominalisierungen problematisch bleibt. In: Probleme und Fortschritte der Transformationsgrammatik (ed. D. Wunderlich), München 1971, S. 189-218.
Yang, S. (Search)
A Search and Data Structure for an Efficient Information System. Preprint Nr. 51 der International Conference on Computational Linguistica 1969 in Sanga Säby bei Stockholm.
Zemanek, H. (Automaten)
Lernende Automaten. In: K. Steinbuch2(ed.), Taschenbuch der Nachrichtenverarbeitung, Berlin 1967., 5. 1383-1450.
Zepic, S. (Nominalkomposita)
Zum Problem der automatischen Erzeugung der deutschen Nominalkomposita. In: Linguistische Berichte 2 (1969) S. 14-24.
Zimmermann, H.H. (Mehrdeutigkeiten)
Zur Auflösung von Mehrdeutigkeiten bei einer maschinellen Analyse des Deutschen. In: Beiträge zur Linguistik 21 (1971) 5. 36 – 49.

1 Ziel und Methode

1.1 Linguistik und elektronische Datenverarbeitung 1/l

Für den Linguisten ist der Computer zu einem nicht mehr ungewöhnlichen Hilfsmittel geworden. Zwei seiner spezifischen Eigenschaften – hohe Verarbeitungsgeschwindigkeit und große Speicherkapazität – haben ihn auf dem Gebiet der Sprachstatistik (Konkordanzen, Indices, Häufigkeitszählungen verschiedenster Art ...) schon lange unentbehrlich werden lassen. Wie verbreitet die Verwendung des Computers gerade in diesem Bereich geworden ist, vermitteln die zahlreichen, seit 1966 in der Zeitschrift Computers and the Humanities (CHum) erschienenen Projektanzeigen. Während dabei aber eher linguistische Randprobleme behandelt werden, ist es ein in wissenschaftlicher Hinsicht bedeutend wichtigeres Vorhaben, den Computer andererseits auch bei der Lösung grundlegender sprachlicher Probleme zu verwenden. Das klassische Beispiel dafür ist die Maschinelle Sprachübersetzung (MT): Bis Mitte der 60-er Jahre gab es vorwiegend praxisnahe Versuche, natürliche Sprache maschinell zu übersetzen; die spezifischen Sprachprobleme wurden eher oberflächlich behandelt.1/2 Mit dem letztlichen Scheitern dieser Ansätze setzte sich zugleich die Ansicht durch, daß eine erste Voraussetzung zur automatischen Sprachübersetzung (und zur Sprachanalyse überhaupt) eine präzise Kenntnis des Sprachsystems ist. Dieser Umschwung ist einmal bedingt durch den 1966 erschienenen Bericht der ALPAC-Kommission 1/3, der einer maschinellen Übersetzung vorwiegend aus ökonomischen Gründen keine Zukunft voraussagte. Eine Reihe unbefriedigend gelöster linguistischer Probleme –

<2> etwa die Auflösung von Mehrdeutigkeiten – ließ daneben eine zeitweilige Zurückstellung von Übersetzungsprojekten sinnvoll erscheinen. 2/1

Neben dieser Forderung nach einer eingehenderen Erfassung des Sprachsystems wird häufig die Frage nach der Formalisierbarkeit sprachlicher Regeln gestellt. In der Linguistischen Datenverarbeitung (LDV) steht dieses Problem notgedrungen mit im Vordergrund, da man in der Verarbeitungsphase auf die formale Sprache eines Rechners angewiesen ist. Daher muß natürlich auf irgendeiner Bearbeitungsstufe eine formale Beschreibung der sprachlichen Regeln (letztlich in einer Computersprache) vorliegen.2/2 Dennoch ist zu betonen, daß die äußere Form der Beschreibung linguistischer Gegebenheiten grundsätzlich beliebig ist; die Brauchbarkeit für die LDV ist nicht zu verwechseln mit dem Grad der Formalisierung (im Sinne einer Mathematisierung). Auch hier können die Anfänge der MT als Beispiel dienen: Die verwendeten linguistischen Gegebenheiten waren hinreichend formalisiert, aber das Sprachsystem damit unzureichend beschrieben.

Viel wichtiger ist der Adäquatheitsgrad der Beschreibung. Er ist in der LDV von entscheidender Bedeutung. Unzureichende oder unexakte Beschreibung führt zu mangelhaften oder fehlerhaften Ergebnissen. Was dabei 'adäquat' bzw. 'unzureichend', 'unexakt' bedeuten, kann nur im Hinblick auf die speziellen Erfordernisse der LDV beurteilt werden: Ein Deutscher beispielsweise, der einen französischen Satz mit dem Verb <S. 3> CHANTER bilden soll, würde wohl kaum sagen LA TABLE CHANTE. Er weiß eben, daß ein Tisch nicht singen kann. Eine Grammatik für Ausländer zum Erlernen des Französischen könnte also – und das geschieht ja auch – weitgehend auf eine semantische Subkategorisierung verzichten und erfüllte dennoch völlig ihren Zweck. Anders dagegen bei einem Computer als Kommunikationspartner, dem dieses enzyklopädische Wissen fehlt: Hier müßte ein komplexeres Kategoriensystem erstellt werden, das solche Sätze wie LA TABLE CHANTE entweder nicht erzeugen oder als (graduell) ungrammatisch erkennen würde. Man hat folglich bei der Diskussion der Adäquatheit zu unterscheiden zwischen gewissen allgemeinen Ansprüchen, die jede Grammatik erfüllen soll, und irgend welchen speziellen Forderungen, die vom jeweiligen Verwendungszweck abhängen. Was die allgemeinen Anforderungen angeht, so werde ich mich im wesentlichen an der Theorie der generativen tranaformationellen Grammatik (TG) orientieren, in deren Rahmen Adäquatheitsfragen einen breiten Raum einnehmen. Die theoretischen Grundlagen der TG 3/1 bilden daher eine wesentliche Basis der vorliegenden Untersuchung.

1.2 Abgrenzung des Themas

Ziel der Grammatik ist es, die Kompetenz eines idealen Sprecher/Hörers zu beschreiben, beliebige Sätze zu bilden und zu verstehen. Ein reales Grammatikmodell ist immer (mehr oder minder) unvollständig, es versucht die Kompetenz eines idealen Sprecher/Hörers nur zu approximieren. Die Grammatik, um die es hier geht, hat die Form eines Computerprogramms. Eine derartige Grammatik ist selbstverständlich <4> gleichfalls unvollständig, insbesondere, weil neben den rein theoretischen auch technische Probleme auftreten.

Die Probleme, die sich aus diesem Phänomen der Diskrepanz zwischen der (theoretischen) Sprachkompetenz des idealen Sprecher/Hörers – und damit etwa der Forderung nach einer völlig adäquaten Grammatik – einerseits und der Sprachkompetenz eines realen Sprecher/Hörers andererseits (etwa der Grammatik des Herrn Meier oder auch des Computers Electrologica X1 4/1) ergeben, sollen hier

für einen Teilbereich der Grammatik, nämlich das Lexikon, behandelt werden. Ich beschränke mich dabei auf den syntaktischmorphologischen Aspekt des Lexikons, obgleich ich mir der Problematik bewußt bin, die Syntax von der Semantik abzugrenzen. Die allgemeinen Regeln, die für ein maschinelles Basislexikon formuliert werden sollen, sind von diesem Problem weitgehend unberührt.

Um die entwickelten theoretischen Vorstellungen zu festigen, ist dieser Arbeit ein anwendungsorientierter Teil angefügt. Er steht in Zusammenhang mit einem Forschungsprojekt, das die Möglichkeiten einer maschinellen syntaktischen Analyse beliebiger deutscher Sätze der Gegenwart erproben sollte.2 In diesem Projekt war jedoch vorausgesetzt, daß die Textformen mithilfe eines vollständigen maschinellen Lexikons vor der Analyse bereits präklassifiziert werden konnten, d.h. daß ihre im Satz möglichen (syntaktischen) Funktionen erkannt sind. In Ergänzung dazu wird in dieser Untersuchung eine Methode zur automatischen Klassifikation 'unbekannter', d.h. mittels des Lexikons nicht präklassifizierbarer Wörter entwickelt und schließlich ein Modell zur automatischen Wörterbucherweiterung für den syntaktischmorphologischen Bereich vorgelegt. Dabei stütze ich mich teilweise auf Erfahrungen, die bei diesem Projekt gesammelt wurden. Die deutsche (Gegenwarts-) Sprache dient zugleich als Materialgrundlage. Einige Teilergebnisse werden also sicherlich nur für die Grammatik der deutschen (oder einer im entsprechenden Fall gleichstrukturierten) Sprache gültig sein, im allgemeinen sind die Ergebnisse jedoch universal anwendbar.

1.3 Vorgehensweise

Den Ausgangspunkt (Kap. 2) der Untersuchung bildet die Betrachtung der Funktion und des Aufbaus eines Lexikons, die sich an der Behandlung der Lexikoneintragung in der TG orientiert. Struktur und Verwendung des Lexikons in einer maschinellen (syntaktischen) Analyse werden anschliessend an den theoretischen Forderungen gemessen, die ihrerseits (Kap. 3) auf ihre Effizienz für die praktische Anwendung hin untersucht werden. Fragen des Adäquatheitsgrades, des Problems von Kompetenz und Performanz und schließlich die Möglichkeit, eine automatische Wortklassifikation und Wörterbucherweiterung in ein maschinelles Analysesystem einzubeziehen, stehen dabei im Mittelpunkt. Den Abschluß des Hauptteils (Kap. 4) bilden schließlich die Beschreibung eines einfachen Modells zur Klassifikation unbekannter Wörter und zur automatischen Lexikonerweiterung, das in das bestehende Saarbrücker Verfahren zur maschinellen Analyse beliebiger deutscher Sätze integriert ist, sowie eine Auswertung entsprechender Tests.

2 Funktion und Aufbau des Lexikons

Im Grunde besteht jedes Lexikon aus einer Reihe von Paaren (Aj,Bi), die gewisse Informationen enthalten. Man kann dabei eine Komponente, also Ai oder Bi, allgemein als Adresse für die andere Komponente auffassen, also für den Ort, wo weitere Informationen zu einem bestimmten <S. 6> Zeichen- oder Symbolkomplex aufbewahrt werden. Entscheidend ist dabei der Aspekt der Informationszuordnung„ wobei die Frage der Anordnung von sekundärer Bedeutung ist. Aus ökonomischen Gründen ist das Lexikon in der Regel geordnet, sei es nach inhaltlichen öder äußeren (etwa alphabetischen) Gesichtpunkten. Darüberhinaus liefert ein Lexikon nur Informationen zu einem bestimmten Sach- oder Inhaltsbereich (Ich betrachte im folgenden die Komponente Ai zugleich als Adresse für die zuzuordnenden Informationen). Etwa wird man in einem Englisch-Deutschen Wörterbuch bei der englischen Eintragung ELONGATE das deutsche Wort VERLÄNGERN finden; in

einem Aussprachewörterbuch führt die Graphemschriftkomponente hin zur Phonemschriftkomponente. Die zuzuordnende Komponente Bi muß allerdings nicht notwendig nur eine Information enthalten. Etwa kann in einem Aussprachewörterbuch aus Redundanzgründen die Adresskomponente Ai (Beispiel ÜBERSCHLAGEN) nur einmal aufgeführt sein, obgleich mehrere Aussprachemöglichkeiten (hier Betonungsunterschiede bei gleichzeitiger Homonymie) unter Bi verzeichnet werden. Ein etymologisches Lexikon enthält als Adresskomponente Wörter, deren sprachliche Entwicklung in der Komponente Bi erklärt wird. Man wird in einem deutschen etymologischen Lexikon also vergeblich die russischen ttbersetzungsäquivalente eines Wortes suchen. Auch der Aufbau führt zu gewissen Konsequenzen: Zwar wird man in einem etymologischen Wörterbuch des heutigen Deutsch – wenn auch mit einiger Mühe – die Entwicklung vieler ahd. Wörter verfolgen können, doch sind sie in jedem Falle ungeordnet und daher nur durch Zufall oder durch Erfassen aller B-Kömponenten aufzufinden, während die Etymölogien nhd. Wörter über die A-Komponenten, also ihre 'Adressen', direkt (oder zumindest – etwa bei Alphabetordnung durch geeignete Intervallsuchverfahren – leichter und schneller) zu erhalten sind. Anders ausgedrückt: Es ist wenig sinnvoll, in einem etymologischen Wörterbuch der deutschen Gegenwartssprache die Entwicklung eines ahd. Wortes, das nur in seiner ahd. Form bekannt ist, verfolgen zu wollen. Daß Inhalt und Aufbau eines Lexikons <S. 7> von seinem Verwendungszweck her wesentlich bestimmt sind, darf damit als eine allgemeine Eigenschaft eines Lexikons angesehen werden.

Ein Lexikon, das bei der Beschreibung der *Sprachstruktur* verwendet wird, muß Informationen enthalten, die für diese Beschreibung relevant sind, und ist somit ein Teil des grammatischen Systems. Eine Grammatik hat zum Ziel, Strukturen einer Sprache zu beschreiben. Die Forderung:

(1) Eine völlig adäquate Grammatik einer Sprache muß jeden beliebigen Satz dieser Sprache so beschreiben, wie er von 'einem idealen Sprecher/Hörer verstanden' 7/l wird

bildet eine Grundlage der generativen Grammatik.

2.1 Die Behandlung des Lexikons in der Theorie der TG 7/2

Die in (1) aufgestellte Forderung kann in ihrer Absolutheit wohl nicht realisiert werden. Allerdings ist sie als Maßstab zu werten für alle (auch transformationelle) Teil-Grammatiken, und es ist ein Hauptziel der TG, sich dieser deskriptiven Adäquatheit zur Sprachkompetenz des idealen Sprecher/Hörers soweit als möglich asymptotisch zu nähern. Dieser Anspruch besteht sowohl für die Ersetzungsund Transformationsregeln wie für die Lexikoneintragungen.

Die Funktion des Lexikons in der TG läßt sich etwa folgendermaßen umgrenzen:

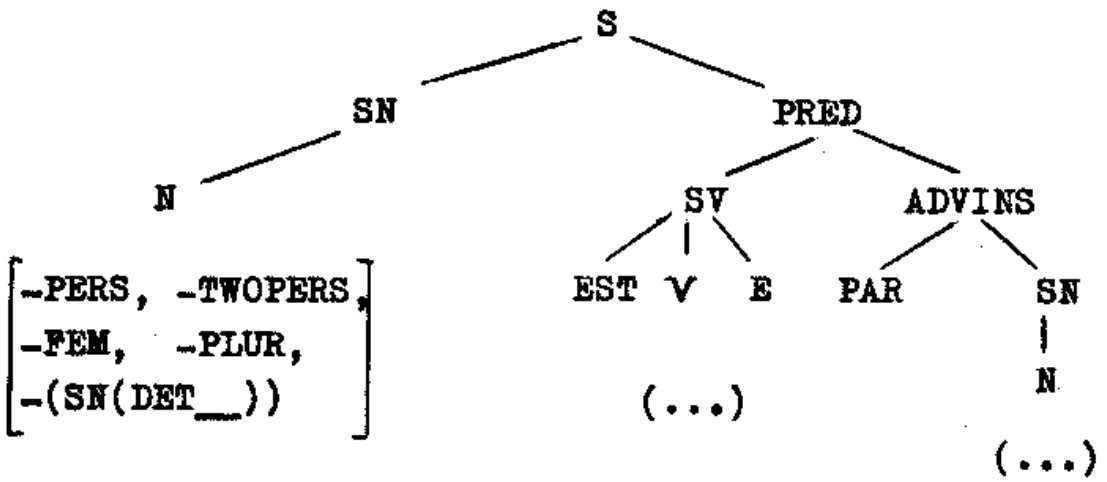
(2) 'Die lexikalischen Eintragungen umfassen ... die Gesamtmenge von Irregularitäten einer Sprache.' 7/3

Damit ist hier also eine Zweiteilung des Sprachsysteme in Regelsystem (Regularitäten enthaltend) und Lexikon ('Irregularitäten enthaltend) vorgenommen. <S. 8>

Ehe auf die Problematik dieser Definition Chomskys jedoch näher eingegangen werden kann, soll zunächst noch der Kopplungsmechanismus beschrieben werden,

über den Lexikon und Regelsystem verbunden werden. Dazu sei der Generierungsprozeß eines Satzes in der TG betrachtet:

Beim Derivationsvorgang wird nach mehreren, hier nicht näher interessierenden Schritten schließlich eine Phase erreicht, in der eine Kette (oder Teile einer Kette) nicht weiter allein mittels des Regelsystems ersetzt oder transformiert werden können. Diese 'präterminale' Kette enthält dann eine Anzahl grammatischer Formative (z.B. Symbole für Perfekt, Possessiv,...) mit dazugehörigen komplexen Symbolen, bestehend aus Kategoriensymbolen wie N (Substantiv) oder Det (Artikel) und spezifizierten Merkmalen wie /+belebt/, /-menschlich/ 8/1. Dazu ein Beispiel von Joyce Friedman, das hier besonders geeignet erscheint, da es aus einem Computer-Experiment entnommen ist: 8/2 Generiert wird der Satz ALCESTE EST REGARDEE PAR CELIMENE.



Für die Darstellung der Lexikonzuordnung sei das erste N des Satzes ausgewählt: Unter den Eintragungen des dem Computer zur Verfügung stehenden Lexikons 8/3 finden sich 10 Belege, die das Merkmal [+N] aufweisen, also für diesen Fall noch in Frage kommen: TRUDEAU, JEAN, ALCESTE, <S. 9> DEGAULLE, CELIMENE, LIVRE, FILLE, PROFESSEUR, MOI, und TOI. Aufgrund des ersten Subkategorisierungsmerkmale PERS (" Pronomen), das negativ spezifiziert ist, müssen die hier im Lexikon positiv spezifizierten Wörter MOl und TOI ausgeschieden werden. Im weiter verlangten Merkmal -TWOPERS stimmen die noch verbleibenden Wörter überein (hier waren die bereits ausgeschiedenen Wörter MOI und TOI voneinander unterschieden). Das nächste Merkmal [-FEM] verlangt ein maskulines Nomen: es entfallen also CELIMENE und FILLE. (Von den restlichen könnte noch PROFESSEUR aufgrund seiner Spezifikation [+NCOM] ausgeschieden werden, wenn im Satz ein Eigenname ([-NCOM]) verlangt worden wäre.) Zur Verfügung stehen also noch (gleichwertig): TRUDEAU, JEAN, ALCESTE, PROFESSEUR und DEGAULLE, wovon durch Zufallentscheidung einer gewählt wird 9/1 (im Satzbeispiel war es ALCESTE). Zu dem gleichen Ergebnis wäre man auch gekommen, wenn man bei einem anderen Merkmal begonnen hätte. Entscheidend ist, daß der Lexikoneintrag in allen Merkmalen die gleiche Spezifizierung aufweisen muß wie die Merkmalkette im Satz. Notwendig ist also eine völlige Übereinstimmung zwischen den Merkmalen des komplexen Symbols und den Merkmalen der Lexikoneintragung sowie die Übereinstimmung mit den Transformationsregeln, die der entsprechenden Lexikoneintragung durch die kontextsensitiven Merkmale mitgegeben sind. 9/2

2.1.1 Der phonologische Aspekt der Lexikoneintragung

Die Menge der phonologischen (bei geschriebener Sprache: graphematischen) Merkmale einer Lexikoneintragung kann im Lexikon der TG in Form einer Matrix aufbewahrt werden. Sie steht dabei zu den übrigen Merkmalen in einem Ist-ein-Verhältnis.9/3 Beispielsweise ist

(3) (boy, [+N, +menschlich ...]) zu lesen: <S. 10>

'boy' kann für das komplexe Symbol [+N, +menschlich ..] eingesetzt werden.

Diese Ersetzungsregel ist ein Teil des Derivationssystems. Das komplexe Symbol [+N, +menschlich] der präterminalen Kette wird in der sog. Terminalen Kette durch 'boy' ersetzt, da die betreffenden Merkmale des komplexen Symbols mit denen des Lexikoneintrags übereinstimmen.

In den 'Aspects' 10/1 bezeichnet Chomsky die Matrix mit D und die übrigen Merkmale mit dem komplexen Symbol C. Wir treffen also hier ebenfalls auf die Zweiteilung der Lexikoneintragung, wobei die Anordnung von der Richtung des Prozesses abhängt, die wir mit 'Derivation' (Erzeugungsprozeß) bzw. 'Analyse' (Erkennungspozeß) bezeichnen wollen. Bei der Derivation entspricht die Komponente Ai dem Symbol C und die Komponente Bi.der Matrix D, bei der Analyse ist es umgekehrt.

Auch die Flexionsparadigmen herkömmlicher Grammatiken lassen sich in der TG in Form von Merkmalen beschreiben. Etwa stehe in der präterminalen Kette ein komplexes Symbol [+N, …, -SINGULAR, +GENITIV, +MASKULIN, DEKLINATIONSKLASSE 1], das Lexikon enthalte den phonologischen Eintrag BRUDER sowie die Merkmale [+N, ..., +MASKULIN, DEKLINATIONSKLASSE 1, UL=3]. Bei der Ersetzung des komplexen Symbols durch den passenden Lexikoneintrag müssen nun zusätzlich die Flexionszeichen angefügt werden. Dies geschieht mittels einer interpretativen phonologischen Ersetzungsregel, bei der auch – wie im obigen Beispiel nötig – die Umwandlung eines Vokals erfolgen kann (Das kann z.B. durch die Verwendung eines Merkmals UL – wie oben angegeben – geschehen, bei dem die Graphemstelle angegeben wird, an der sich der umzulautende Vokal befindet). Die beiden Merkmale [-SINGULAR, +GENITIV] des komplexen Symbols müssen im Lexikon nicht <S. 11> explizit verzeichnet sein, da sie anhand der Deklinationsklasse erzeugt werden können. Sie sind also redundant und werden durch entsprechende (Redundanz-)Regeln erfaßt.11/1 Anders ausgedrückt: Diese letztgenannten Merkmale werden nicht mit den Lexikonmerkmalen explizit auf Übereinstimmung hin überprüft, sondern dienen nur der endgültigen Ausformung des Wortlauts.11/2

2.1.2 Die Merkmalklassifikation

Es würde zu weit führen, hier mehr als einen kurzen Abriß des Ausbaus einer generativen Grammatik im Hinblick auf die Funktion der Lexikoneintragungen, besonders der syntaktischen Angaben, zu geben. Wir wollen Einzelheiten – wie auch bei der Beschreibung der phonologischen Komponente geschehen – vermeiden und nur die Voraussetzungen referieren, die zum Verständnis der allgemeinen Grundlagen eines Lexikons für notwendig erachtet werden.

Eine Grundfrage der TG ist die nach der Grammatikalität eines sprachlichen Ausdrucks, wobei 'ungrammatisch' etwa als 'Abweichung von einer (wohlgeformten) Grammatikregel' paraphrasiert werden könnte. Ohne auf die postulierten verschiedenen Grammatikalitätsgrade hier einzugehen3, sei zunächst angenommen, daß etwa ein Sprecher nur solche wohlgeformten Sätze sprechen soll. Neben einem korrekten Regelsystem muß er auch über ein exakt spezifiziertes Lexikon verfügen, um eine richtige Auswahl aus der Menge der Lexikoneintragungen vornehmen zu können. Diese Selektion erfolgt – wie schon bei der Erläuterung des Zuordnungsprinzips dargestellt – anhand von Merkmalen, die über die Verwendungsmöglichkeit der phonologischen Komponente Auskunft geben. <S. 12>

(a) Eine syntaktische Teilkomponente enthält Angaben zu grammatischen Kategorien und Subkategorien sowie zur kontexabhängigen syntaktischen Selektion;
(b) eine transformationelle Teilkomponente gibt die beim Anwenden von Transformationsregeln bei diesem Lexikoneintrag beachtet werden müssen;
(c) die semantische Teilkomponente soll schließlich die Verwendung der phonologischen Matrix in Abhängigkeit von der semantischen Struktur des Kontexts weiter determinieren.

Bei einem Verstoß gegen (a) oder einem entsprechend unrichtigen Lexikoneintrag könnten etwa ungrammatische Sätze folgender Art entstehen:

Er BRUDERt heute. (Falsche Wortklasse N -* V)
Er unterliegt deN Bruder. (Falsche Subkategorisierung Dativ --s Akkusativ)
Das Wasser fließt IHN. (Fehlende oder falsche Rektionsoder Selektionsangabe)

Falls etwa ein Lexikoneintrag in bezug auf (b) falsch oder unvollständig spezifiziert ist, kann etwa aus dem Satz

ER HOERT MICH.

die fälsche Reflexivierungstransformation

ER HOERT MICH SELBST.

entstehen, wenn eine Referenzidentität zwischen ER und MICH hergestellt wird, wobei MICH falsch oder unvollkommen klassifiziert ist: Eine derartige Transformation hätte nur bei 'Ich höre mich (- selbst), oder 'Er hört sich (-->selbst), erfolgen dürfen.12/1

Die semantische Klassifikation (c) soll verhindern, daß Sätze wie '*Die Katze frißt den Elefanten.' als Normalsätze entstehen oder betrachtet werden.12/2 <S. 13>

Die hier erwähnten Klassifizierungsmöglichkeiten sollten nur einen Einblick in das mögliche Inventar der Lexikonmerkmale geben und sind bei weitem nicht vollständig. Es bleibt nachzutragen, daß die TG dazu ein eingehend formalisiertes Beschreibungssystem anbietet und der maschinellen Bearbeitung dadurch entgegenkommt.

2.1.3 Das Prinzip der Vereinfachung (Redundanz)

Ein weiterer Aspekt, der bei der Erstellung eines Lexikons zur maschinellen Analyse von Sätzen interessant und wichtig erscheint, wird in der TG unter dem Begriff der 'Redundanz' aufgezeigt, den wir hier in bezug auf die Lexikoneintragung betrachten wollen. In den 'Aspects' 13/1 werden Möglichkeiten zur einfacheren Darstellung von Lexikon-Merkmalen untersucht. Das Ergebnis kann man etwa auf die folgende einfache Formel bringen:

> (4) Immer dann, wenn sich lexikalische Merkmale hierarchisch derartig gliedern lassen, daß ein Merkmal eine Teilmenge zu einem in der Hierarchie höherstehenden Merkmal darstellt, verkörpert dieses Merkmal implizit auch diese allgemeinere Merkmalspezifikation. Es genügt in diesem Falle also, nur das letzte (niedrigste) Element im Lexikon zu vermerken.

Etwa repräsentiert das Merkmal [+menschlich] zugleich das allgemeinere Kennzeichen [+belebt]. 13/2 Dies gilt ebenso für Subkategorisierungsangaben wie für kontextsensitive Selektionsmerkmale, die alle unter den Begriff der syntaktischen Redundanz gefaßt werden. Dieses Phänomen läßt sich auch bei der phonologischen Komponente der Lexikoneintragung beobachten. Auch hier sind allgemeine regelmäßige Eigenschaften festzustellen. Die phonologische und die syntaktische Komponente des Lexikons lassen sich damit durch Formulierung geeigneter Redundanzregeln quantitativ einschränken. Von Chomsky wird dies so zusammengefaßt: <S. 14>

> (5) 'Die Redundanzregeln ... konstatieren allgemeine Eigenschaften ... und machen daher Merkmalspezifikationen in Lexikon-Einheiten da überflüssig, wo diese nicht idiosynkratisch sind.' 14/1

Der Maßstab, mit dem die lexikoneintragung hier gemessen wird, ist ökonomischer Natur. Daraus läßt sich folgern, daß sich die Verwendung von Redundanzregeln in der Grammatik ebenfalls diesem Grundsatz der Vereinfachung eines Grammatiksystems stellen muß. 14/2 Obgleich nämlich prinzipiell zugestanden werden muß, daß redundante Eigenschaften bei Lexikonmerkmalen eliminiert und durch eine Regel erfaßt werden sollten, lassen sich Fälle denken, in denen eine Redundanzregel ökonomisch nicht effektiv ist, da sie ihrerseits den Analyse- oder Syntheseprozeß verkompliziert.

Wir nehmen als Beispiel die sog. 'starken' Verben im Deutschen und betrachten dabei die morphologische Komponente. Je nach Art der Abwandlung eines Stammvokals (evtl. auch der Art des Konsonantenwechseis) lassen sich diese Verben in Gruppen (nicht zu verwechseln mit den wohldefinierten historischen Ablautreihen) zusammenfassen. Einigen dieser Gruppen kann man viele Verben zuordnen; beispielsweise verzeichnet Dietrich 14/3 unter dem Typ 11.1.1 (Vokalwechsel i – a – ae – u, Beispiel BINDEN- BAND – BAENDE – GEBUNDEN) eine ganze Reihe: DINGEN, DRINGEN, GELINGEN, KLINGEN, RINGEN, SCHLINGEN, SCHWINGEN, SCHWINDEN, SINGEN, SINKEN, SPRINGEN, STINKEN, TRINKEN, WINDEN, WRINGEN, ZWINGEN, MISSLINGEN; unter dem Typ 4.2.1 lassen sich einige wenige einordnen – wie BRATEN (BRATE – BRAET – BRIETE) werden noch HALTEN, RATEN, SCHLAFEN, BLASEN gebildet; für manch eine 'Gruppe' gibt es nur ein einziges zugehöriges <S. 15> Simplex-Verb, z.B. FALLEN, FANGEN, ESSEN. Die 193 starken oder

unregelmäßigen Verben (Simplizia) des Deutschen, die Dietrich 15/1 aufführt, sind aufgrund der zur Bildung aller Flexionsformen nötigen Regeln von ihm in 44 Hauptklassen eingeteilt worden. Berücksichtigt man die Vokal- und sonstigen Merkmalwechsel, ergeben sich sogar 68 Subklassen, wobei 40 dieser Subklassen jeweils nur ein einziges Verb zugeordnet worden ist. Billmeier 15/2 kommt in seiner Einteilung bei 196 'Formativen' (˜ beim Verb die morphologische Komponente des Lexikoneintrags eines Simplex) sogar auf 84 'Flektionsklassen', wobei z.B. als Formative für ZIEHEN etwa nur noch 'Z-', für GIEßEN nur noch 'G-' eingetragen werden; die übrigen Grapheme einer Flexionsform werden anhand von Flexionsklassenmerkmalen mittels eines entsprechenden Algorithmus erzeugt bzw. erkannt. Ähnlich wie bei Dietrich sind hier die Verben SEIN, WERDEN und TUN gesondert behandelt. Dietrich führt für diese und einige weitere Hilfsverben z.T. weitere Flexionsformenmerkmale ein 15/3, Billmeier behandelt sie gänzlich gesondert. Dies ist vor allem bei SEIN nötig, da z.B. bei den Flexionsformen SIND/WAR überhaupt keine graphematischen Übereinstimmungen vorhanden sind, als gemeinsames Formativ also '-' eingesetzt werden müßte.

Während bei Dietrich die linguistisch adäquate Beschreibung – oder die Suche nach ihr – im Vordergrund steht (unabhängig von einer evtl. möglichen Umsetzung in einen maschinellen Algorithmus), argumentiert Billmeier auch äuf der Basis der günstigeren Umsetzung in ein (sog. interaktives, d.h. dialogartiges) Computersystem: Die Speicherung der Formative benötige weniger Platz im Lexikon als die Speicherung von Wortformen oder mehreren Worteinträgen bei einem einzigen Verb. Der höhere Aufwand an Regeln erscheint ihm angemessen. An einem Beispiel konkretisiert heißt dies, daß etwa bei der Erzeugung oder Erkennung der Wortform ZIEH bei Billmeier (aufgrund des Konsonantenwechsels H – G <S. 16> im Imperfekt) die Graphemfolge IEH mit dem Formativ 'Z-' verknüpft werden muß; IEH zumindest steht also in einer Liste des Regelteils. Die konsequente Verkürzung des Lexikons zieht also eine teilweise idiosynkratische Erweiterung des Regelteils nach sich. Im Hinblick auf die Präfigierungsmöglichkeiten der starken Verben scheint eine Einschränkung der Anzahl der morphologischen Komponenten eines Verb-Lexikons hier durchaus angebracht. Es steht außer Frage, daß es effektiv ist, die Verbgruppen mit vielen Elementen mittels bestimmter, im Lexikon zu verzeichnender graphematiseher Merkmale zu erkennen und zu bearbeiten. Problematisch ist dies nur, falls das Schema auch bei Einzelfällen verwendet wird (obgleich dem grundsätzlich nichts entgegensteht), da das Regelsystem dann mit eben diesen idiosynkratischen Regeln belastet wird, die ein einziges oder wenige besondere Verben spezifizieren. Es bleibt der Erfahrung zu überlassen, ob in den Grenzbereichen, wie sie auch im System der starken Verben auftreten, eine (redundante) Lexikoneintragung oder eine (Redundanz-)Regel effektiver ist.

Ganz allgemein scheint zu gelten – und dies zu erarbeiten war der eigentliche Sinn dieses kleinen Exkurses -, daß die Redundanzregel zwar den Ablauf des Verstehensprozesses, aber nicht notwendig das Ergebnis beeinflußt, sondern nur eine Verlagerung des 'Wissens' von der Regel ins Lexikon oder umgekehrt nach sich zieht. Begreift man Lexikon und Regelsystem mehr als eine Einheit denn als zwei scharf zu trennende Komponenten des Sprachsystems 16/l, so fällt es nicht schwer, diesem Redundanzproblem eher eine ökonomische als eine primär-linguistische Bedeutung zuzuschreiben. Bei der Umsetzung des Grammatik-Lexikon-Systems für einen Parser wird das Problem der technischen Realisierung mitentscheidend sein. Sieht man von dem nicht immer klar abzuwägenden Faktor der benötigten Rechenzeit einmal ab, so <S. 17> ist folgendes zu beachten: Ein Lexikoneintrag ist technisch einfacher zu erstellen als eine komplexe Redundanzregel im Algorithmus;

ähnliches gilt für die Bearbeitung, da bei jedem wiederholten Auftreten einer (sonst mit allen Besonderheiten verzeichneten Sonderform) das Regelsystem dieses 'Wissen' auch stets neu erarbeiten muß. Ein stärker redundanter Lexikonanteil kann durchaus zu einer Entlastung des Regelteils eines Parsers führen. 17/1

2.1.4 Wortbildung und Produktivität

In engem Zusammenhang mit der Frage der Vereinfachung des Lexikons steht das Problem der Wortbildung. In der herkömmlichen Terminölogie lassen sich im Wesentlichen zwei Wortbildungsprozesse unterscheiden: die Komposition und die Ableitung. Unter Komposition versteht man allgemein die Zusammensetzung von zwei oder mehreren selbständigen Wörtern (freien Morphemen) zu einem neuen Wort 17/2; Ableitung ist die Bildung eines neuen Wortes aus einem selbständigen Wort (freien Morphem) mithilfe von unselbständigen Affixen (gebundenen Morphemen). Es ist nun zu fragen, inwieweit sich diese Wortbildungsprozesse formalisieren lassen. Für die TG ist die Lösung dieses Problems von entscheidender Bedeutung, da im Grammatiksystem Regeln zur Komposition und Ableitung vorgesehen sein müssen, die zugleich die Forderung zu erfüllen haben, daß nur 'grammatische' Wörter gebildet werden. Eine 'Produktivität' von Wortbildungsregeln im traditionellen Sinn (etwa im Hinblick darauf, daß gewisse Ableitungen auf -bar oder -heit im 20. Jh. besonders 'produktiv' sind, da mit ihrer Hilfe heute viele Wörter abgeleitet werden) kennt die TG nicht, zumindest nicht in bezug auf die deskriptive Seite der Grammatik: in Hinsicht auf die 'Performanz', also die Realisierung einer <S. 18> entsprechenden Regel, decken sich allerdings die Begriffe wieder. In der TG wird unter 'produktiv' (für den Bereich der Kompetenz) vielmehr verstanden, daß eine Regel *an sich* gewisse Strukturen (oder Wörter) erzeugt. Sie muß so abgefaßt sein, daß sie anhand von Restriktionsmerkmalen nur grammatische Strukturen oder Wörter zuläßt. Anders ausgedrückt: Ein Lexikoneintrag (freies Morphem) muß Hinweise darauf enthalten, ob eine bestimmte Wortbildungsregel angewendet werden kann oder nicht, also eine Regel in Hinsicht auf diese Bildungsmöglichkeit produktiv ist oder nicht. Beispielsweise müßte das freie Morphem STOER etwa das Merkmal 18/1 'UNG-Ableitung möglich' besitzen. Bei der Vielzahl der möglichen Ableitungen ergäben sich lange Listen von Merkmalangaben bei einem einzigen Morphem, etwa +BAR, +BARKEIT, +UNG, -HEIT, -KEIT, +ER, ..., sofern diese negativen und positiven Spezifizierungen nicht aus Redundanzgründen entfallen können.

Ein derartig abgeleitetes Wort wird in der TG durch Transformationsregeln erklärt. Dabei ist also nur die Oberflächenstruktur verändert. Das Wort STOERUNG ist etwa ein 'defektives Prädikat',dem die gleiche Tiefenstruktur zugrundeliegt wie dem Satz JEMAND STOERT ETWAS. Chomsky hat dieses Problem ausführlicher behandelt.18/2 Auch er kann sich dem Dilemma nicht ganz entziehen, daß die Anwendung von Wortbildungsregeln teilweise sehr starken Restriktionen unterliegt, die kaum oder nur sehr schwer zu formulieren sind. Er versucht dem dadurch zu begegnen, daß er – herkommend von den Ergebnissen Halles, die dieser bei Untersuchungen der phonologischen Komponente gewonnen hatte – drei Möglichkeiten unterscheidet3, die im Lexikon zu <S. 19> vermerken sind: a) vorkommend; b) möglich, aber nicht realisiert; c) nicht möglich. Vor allem bei der Komposition aber (man denke gerade an die im Deutschen möglichen 'Augenblickskomposita') kann eine Prädiktabilität nach dem heutigen Stand der Linguistik nicht immer eindeutig festgelegt werden. Gewissen Wortbildungen beispielweise, etwa Komposita wie ESELSBRÜCKE oder BÜRGERMEISTER, also vor allem solchen mit metaphorischem Charakter, läßt sich heute keine oder nur eine

höchst komplizierte, kaum mehr formalisierbare Tiefenstruktur zuordnen, die auf den Kompositionselementen aufbaut. Vor allem läßt sich die semantische Komponente des Kompositums nicht öder nicht mehr aus den semaitischen Merkmalen der Kompositionsteile erschließen. Derartige Wörter sind also als idiosynkratische Einheiten zu betrachten und direkt als solche ins Lexikon aufzunehmen 19/1. Es ist jedoch – dies darf als allgemeine Forderung der TG angesehen werden – stets zu prüfen, ob sich formalisierbare Strukturen phonologischer, syntaktischer und semantischer Art erkennen lassen, die eine Aufnahme der komplexeren Fügung in das Lexikon überflüssig erscheinen lassen.

Ein in diesem Zusammenhang auftretendes Problem ist das der 'Lücken' im Wortschatz. Während die traditionelle Grammatik mittels der 'Produktivitätskomponente' sog. 'Neubildungen' zuläßt, gilt – wie schon erwähnt – in der TG der Grundsatz, daß das gesamte System – und damit auch das Lexikon – deskriptiv erfaßt sein muß. (Es sind also nur noch Lücken im Sinne von 'vorkommend, aber noch nicht realisiert' möglich.) Ein Ausweg bietet sich also in der Annahme, daß mehr oder minder zufällige Lücken im Wortschatz bestehen, wobei übergenerelle Regeln, die bisher nicht in der Performanz vorhandene (aber dennoch grammatisch <S. 20> richtige) Wörter erzeugen, eine derartige Lücke zu schließen in der Lage sind.20/1

Bei der (menschlichen oder maschinellen) 'Analyse' gegebener sprachlicher Daten stellt sich dieses Problem unter einem anderen Gesichtspunkt dar: Die Wörter sind bereits gebildet. Die Wortgrenzen lassen sich im allgemeinen leicht erkennen (besonders bei schriftlichen Sprachäußerungen: hier gelten Zwischenraum oder Satzzeichen als Wortbegrenzungen). Wenn man beispielsweise eine Reklameschrift liest wie PALETTIEREN AUCH SIE!, so läßt sich noch relativ einfach die verbale Ableitung aus dem Substantiv PALETTE (ein Transport-System der Bundesbahn) nachvollziehen; ähnliches gilt für Wortbildungen wie BARZELN (Eigenname BARZEL). Der Verstehensprozeß ist hier abhängig von der Kenntnis des zur Ableitung verwendeten Lexikoneintrags (PALETTE, BARZEL) und der Kenntnis der Ableitungsregel, evtl. auch vom enzyklopädischen Wissen des Interpretierenden. Ähnliches gilt für die Komposition, die besonders im Deutschen oft keine syntaktischen Zusammenhänge mehr erkennen laßt (HOSENANSICHT – Ansicht, Anblick einer HOSE; VIETNAMKRIEG – Krieg *in* Vietnam) oder gar zu Mehrdeutigkeiten führt, die erst nach der Einbeziehung des Kontexts aufzulösen sind (BERLINREGELUNG = Regelung über Berlin, sprachlich möglich wäre auch: eine Regelung anderer Art, die in Berlin. getroffen wurde). 20/2 Zusätzlich hat eine Überprüfung der grammatischen Wohlgeformtheit zu erfolgen, um fehlerhafte Bildungen (wie SICHERKEIT oder TAPFERHEIT – sprachlich korrekt wäre SICHERHEIT UND TAPFERKEIT) zu ermitteln. Falls ein solches Erkennungssystem vorliegt, so ist damit auch den Ansprüchen der TG genüge getan. Anders gesagt: Eine (maschinelle) Analyse ist solange nicht adäquat, als >S. 21> sie nicht auch Abweichungen von der Grammatikalität registrieren kann,

2.2 Die Verwendung des Lexikons in der Maschinellen Sprachanalyse

Vergleicht man die aufgezeigten theoretischen Anforderungen an ein Lexikon mit den bisherigen Praktiken in der linguistischen Datenverarbeitung, besonders der maschinellen Sprachanalyse, so stellt man in den Resultaten kaum einen größeren Unterschied fest. Dies gilt vor allem für das Problem einer möglichst kompakten, jedwede Redundanz vermeidenden Speicherung und die Frage der günstigsten Gliederung des Wortschatzes. Schon aus technischen und finanziellen Erwägungen

heraus – man denke an die mangelhafte Speicherkapazität der ersten Computergenerationen und an das starke Anwachsen von Rechenzeit bei der Wörterbuchsuche mit zunehmendem Umfang der Lexika – war die Kodierung der lexikoneintragungen – allerdings weniger aufgrund linguistischer Kriterien – konsequent optimiert worden. 21/1 In den Fällen, in denen man aus praktischen Gründen – sei es, daß nur ein beschränktes Wortmaterial als Grundlage diente oder daß es im Zusammenhang mit dem Forschungsziel ohne Belang war – auf eine Vereinfachung des Wörterbuchs verzichtete, wurde etwa auf die Vorläufigkeit dieser Maßnahme 21/2 oder (etwa bei relativ flexionsarmen Sprachen wie dem Englischen) auf die geringere Bedeutung dieses Aspekts 21/3 verwiesen. In allen diesen Fällen liegt die schon erwähnte Überlegung zugrunde, daß bei redundanten Lexikonangaben das Ergebnis der Reduktion oder Expansion das gleiche sein kann wie bei optimaler Ersparung, wo die betreffenden Informationen erst noch anhand von Redundanzregeln ermittelt werden müssen.

Weitaus problematischer als die Frage, in welcher Form <S. 22> und Inhalt dieser Informationen. Bisher existiert noch kein Algorithmus zur Generierung oder Reduktion von Sätzen, der den Ansprüchen der TG in bezug auf die deskriptive Adäquatheit der Grammatik einer natürlichen Sprache qualitativ und quantitativ auch nur nahekommt. 22/1 Die bisherigen (auch computeroperativen) Transformationsgrammatiken stellen entweder nur Teilgrammatiken dar, indem sie z.B. die syntaktisch-morphologische Komponente des Verbs 22/2 oder die der Komposita usw. beschreihen, oder zeichnen letztlich nur den Aufbau einer solchen Grammatik, ohne die aufgezeigten Teile über Beispiele hinaus mit dem Sprachmaterial zu füllen.

Im Zusammenhang mit der maschinellen Sprachanalyse ist also festzustellen, daß ein einem System einer natürlichen Sprache weitgehend adäquater Parser gegenwärtig nicht erstellt werden kann, da es nicht möglich ist, eine entsprechend vollständige Grammatik zugrundezulegen. Genausowenig wie man deshalb der TG jedoch ihren Wert absprechen kann, ist der maschinellen Sprachanalyse – und der maschinellen Sprachübersetzung – aus diesem Grund die Daseinsberechtigung zu entziehen.22/3 Der Aufbau eines Reduktionsalgorithmus für eine Sprache ist – zumindest gegenwärtig – nichts anderes als die Fortsetzung der Bestrebungen, eine möglichst adäquate Grammatik dieser Sprache zu erhälten, mit anderen Methoden und Mitteln. Die maschinelle Sprachanalyse ist also in der Aufbauphase – und das bedeutet eigentlich: <S. 23> *immer* (da eine vollständig adäquate Grammatik nicht realisierbar ist) – in den Prozeß zur Ausarbeitung einer Grammatik integriert. Zugleich stellt sie eine Verifikation der bisherigen Kenntnisse von einer Sprache dar.

Für die Computerlinguistik gilt aber auch, was für die Linguistik allgemein postuliert werden kann. Prinzipiell sind bei der Erschließung des Sprachsystems zwei Vorgehenswege denkbar: Den ersten will ich den Weg des qualitativen Bescheidene nennen. Man wählt ein kleines Teilproblem – etwa die Steigerung der Farbadjektive – aus und betreibt eine Art Grundlagenforschung, etwa auf der Basis von Informantentests oder auch mithilfe von Textuntersuchungen. Das Ergebnis ist eine ziemlich genaue Beschreibung eines kleinen Ausschnitts des Sprachsystems, der in das gesamte System zu integrieren ist, mit dessen Hilfe und auf dessen Grundlage die übrigen Bausteine in Angriff genommen werden können.

Der zweite Weg scheint der unbescheidenere zu sein: die Entwicklung eine alle möglichen Bestandteile umfassenden Grammatiksystems 23/1 – denn ein Parser

stellt nichts anderes dar als eine operative Grammatik – auf Kosten einer unexakteren, oberflächlicheren Beschreibung und damit mit im Einzelfall ungenaueren Ergebnissen. Ich will einmal von dem *anwendungsorientierten* Argument absehen, das hier oft angebracht wird: daß derartige Verfahren nämlich durchaus sinnvoll seien, sofern. sie einigermaßen brauchbare Ergebnisse liefern (also etwa dahingehend, daß eine maschinelle Übersetzung nützlich ist, wenn es gelingt, wissenschaftliche Literatur der Forschung oder der Industrie durch eine schlichte, aber einigermaßen verständliche Rohübersetzung früh- und damit rechtzeitig zugänglich zu machen). Ich bin der Auffassung, daß dieser zweite, auch in Saarbrücken verfolgte Weg über die vordergründige Anwendungsorientiertheit hinaus von hohem heuristischen Wert ist im Hinblick <S. 24> auf eine Vertiefung der Kenntnisse über ein Sprachsystem und die Sprache überhaupt. Wenn eine maschinelle Analyse auf dieser Basis erstellt wird, bietet sich zusätzlich ein weiterer Vorteil: Es entsteht eine Art Rückkopplungseffekt dadurch, daß der Parser die Schwächen eines formalen Systems – in unserem Fall also des Grammatiksystems der Deutschen Gegenwartssprache (und zwar sowohl im Hinblick auf die Lexikoninformation wie in bezug auf den Analysealgorithmus) 'schonungslos' offenlegt. Dieser Weg des Trial and Error führt zu Lerneffekten und damit im ungünstigen Fall zur Änderung des Beschreibungssystems, im günstigeren Fall zur Verfeinerung und damit asymptotischen Näherung an das existierende Sprachsystem.

Ich will diese beiden Wege nicht gegeneinander abwägen, zumal sicherlich methodische Mischformen notwendig und möglich sind. Der erste Weg der Beschränkung auf ein Teilgebiet wäre wohl ständig daran zu messen, inwieweit er den sprachlichen Materialmassen Rechnung trägt – ich denke hier vor allem an die existierenden semantischen Subkategorisierungssysteme 24/1 -; der zweite, auf ein Gesamtsystem als Nahziel ausgerichtete Weg ist der Gefahr einer allzu großen Oberflächlichkeit ausgesetzt.

Hier möchte ich einer Auffassung Chomskys begegnen, die in der Linguistik zu einer Vorliebe der Wissenschaftler für die theoretischen Aspekte und zur Vernachlässigung der pragmatischen Komponente geführt hat. Chomsky argumentiert nämlich, 24/2 daß die sprachlichen Daten evident genug seien und eine weitere 'Befeilung', also eine noch exaktere Beschreibung der Phänomene, wenn derzeit nicht überflüssig, so doch von geringerer Bedeutung sei 'für die Probleme, die anstehen'.24/3 Er verweist auf die ferne Zukunft, in der tieferliegende Fragen dann mithilfe einer intensiveren Durchleuchtung der Daten gelöst werden könnten. Wenn man einmal davon absieht, daß gerade diese <S. 25> Argumentation wesentlich dazu beigetragen hat, eine Vielzahl 'reiner' Linguisten hervorzubringen (die sich oft genug an ihrer eigenen linguistischen Kompetenz genügen lassen) und andererseits diejenigen, die sich vorwiegend mit der Bewältigung großer Datenmengen befassen, leicht pejorativ als 'angewandte' Linguisten zu klassifizieren, so wurde gerade hierdurch zunächst auch die weitere Richtung der TG festgelegt in der Dreiteilung Phonologie – Syntax – (interpretative) Semantik, wobei der Syntax (eben jenem Bereich, der in der Tat nahezu völlig evident ist) die primäre Stellung (Basis) zugeordnet wurde. Erst die generative Semantik (Fillmore, Lakoff u.a.) scheint allmählich hier die entscheidende Wende einzuleiten, eine Wende auch in Richtung einer verfeinerten Beschreibung' der sprachlichen Merkmale in den Lexikoneintragungen.

2.2.1 Struktur und Inhalt der Wörterbucheintragungen

Die lexikoneintragung ist in Aufbau und Inhalt abhängig von dem Gebrauch, der im Analyseprozeß von ihr gemacht werden soll. Je anspruchsvoller dabei der Parser ist, um so differenzierter ist ihr Informationsgehalt. Wenn zui Beispiel Sätze zugelassen sind wie DIE KATZE FRISST DEN ELEFANTEN (falls keine semantischen Restriktionen berücksichtigt werden), kann bei dem Wörterbucheintrag KATZE die Information fehlen, die man verbal etwa formulieren könnte: 'Es können (normalerweise) nur folgende Lebewesen gefressen werden: Mäuse, Ratten, junge Vögel ... '. Ähnliches gilt auch für syntaktische Angaben. Falls ein Parser nicht zwischen präpositionaler Verbergänzung und Umstandsergänzung unterscheiden kann, ist es sinnlos, im Lexikon bei bestimmten Verben (z.B. HOFFEN) etwa das Merkmal 'Mit präpositionaler Ergänzung AUF' anzufügen. Falls kein morphologischer Erkennungsalgorithmus zur Verfügung steht, der etwa aus dem Lexikoneintrag WUEST (Adverb/Adjektiv) in Verbindung mit dem Suffix -LING das Substantiv WUESTLING ableiten kann, ist es nicht nötig, bei WUEST einen Hinweis <S. 26> auf die entsprechende Wortbildungsmöglichkeit anzubringen; stattdessen wäre etwa ein eigener Worteintrag WUESTLING im Lexikon erforderlich.

2.2.1.1 Die morphologische Komponente des Lexikons

Ehe näher auf die mögliche Struktur eines Computer-Lexikons eingegangen wird, seien ansatzweise einmal dem menschlichen Benutzer verfügbare Lexika exemplarisch daraufhin untersucht, inwieweit sie den Forderungen der TG entsprechen, daß möglichst nur Irregularitäten einer Sprache im Lexikon verzeichnet sein sollen. Ich beschränke mich dabei auf einsprachige Wörterbücher und wähle dazu zwei der gegenwärtig 'gängigen' deutschen Wörterbücher aus, den Rechtschreib-DUDEN und das Deutsche Wörterbuch von Wahrig (im Text zitiert als 'Wahrig'); ich beschränke mich weiter auf die Überprüfung der Verwendung morphologischer Redundanzregeln.26/1

1. Stichprobe: Die Behandlung der Eintragungen für starke Verben

Im DUDEN sind die starken Verben nur unter der Grundform (Infinitiv) an der entsprechenden Alphabetstelle aufgeführt. Es wird also vorausgesetzt, daß der Benutzer die Unregelmäßigkeiten in der Konjugation bereits kennt, da er sonst kaum von einem – in irgendeinem Text angetroffenen GEGESSEN zu dem Eintrag ESSEN findet. Dort ist allerdings (was er in diesem Fall teilweise schon weiß) die Besonderheit der Flexion verzeichnet. Anders dagegen im 'Wahrig': hier wird prinzipiell jede unregelmäßig gebildete Form an ihrer Alphabetstelle aufgeführt (also etwa GESUNGEN, SANG, SAENGE als Lexikoneintrag) mit Hinweis auf die <S. 27> Grundform; bei der Grundform wird zusätzlich auf die Tabelle der unregelmäßig konjugierten Verben verwiesen.27/1 Im 'Wahrig' wie im DUDEN werden bei den präfigierten unregelmäßigen Verben die einzelnen 'Stammformen' nicht wieder aufgeführt, sondern auf den Typ (Wahrig') oder auf das Simplex (DUDEN 27/2) verwiesen. Hier wird vom Benutzer also die Anwendung einer 'Redundanzregel' verlangt. Der DUDEN geht diesen Schritt auch noch für die unregelmäßigen Simplicia weiter; er hat keinen eigenen Lexikoneintrag für diese Verben, d.h. die Zuordnung ist nur über eine generative Regel (die dem Benutzer mit allen bei den starken Verben möglichen Speizifikationen bekannt sein muß) oder, was wahrscheinlicher ist, über die lexikalischen Kenntnisse des Benutzers möglich. Im 'Wahrig' werden 'redundante' Lexikoneinträge aufgeführt, die es dem

Benutzer erlauben, ohne Kenntnis der speziellen morphologischen Zuordnungsregel den Zusammenhang zwischen der von der Grundform abweichenden Wortform und dem Hauptstichwort, eben der Grundform, herzustellen, wo weitere Angaben zu dem Eintrag zu finden sind. In beiden Lexika ist – wie in den traditionellen Wörterbüchern üblich – ein Paradigma angegeben, die Flexionsformen sind nach Regeln zu erzeugen, die in einem Vorauskapitel (DUDEN) bzw. in einem Grammatikabriß ('Wahrig') beschrieben sind: auch hier sind also morphologische Redundanzregeln anzuwenden.

2. Stichprobe: Adjektiv-Präfigierung mit UN-

In der Behandlung der UN-Präfigierung finden sich bei beiden Lexika kaum Unterschiede. Ein Test, bei dem alle bei einem deutschsprachigen Dichter 27/3 belegten, mit UN <S. 28> präfigierten Adjektive (und Deverbativa) mit den Eintragungen der beiden Lexika verglichen wurden, zeigte eine auffallende Übereinstimmung: Von den 37 belegten verschiedenen Wörtern waren im DUDEN und im 'Wahrig' jeweils 30 als Stichwörter aufgeführt; in beiden Lexika fehlten übereinstimmend die 6 Belege UNEMPFANGEN, UNERLÖST, UNERMESSEN, UNGEENDET und UNHEILIG; im DUDEN fehlte darüberhinaus UNBEWOHNT, im 'Wahrig UNGEMESSEN!

Trotz dieser augenscheinlichen Übereinstimmung zeigt diese Untersuchung, daß ein verhältnismäßig hoher Anteil an Wörtern (mehr als 20 %)nicht unmittelbar über einen Lexikoneintrag zugeordnet werden kann. Eine morphologische Analyse durch den Benutzer wird in diesen Fällen unumgänglich. Natürlich könnte man annehmen, daß die Lexika – um nicht zu unhandlich zu werden – sich an der Gebrauchshäufigkeit orientierten und nur eine Auswahl der Negationspräfigierungen geben wollen, ohne intensivere Rücksicht auf mögliche Redundanzregeln. Daß dabei subjektive Wertungen eine Rolle spielen, ist wohl offenkundig. Wörter wie UNHEILIG, UNERMESSEN, UNERLÖST und UNBEWOHNT erscheinen mir nicht ungewohnter oder ungebräuchlicher als etwa im 'Wahrig' verzeichnete Wörter wie UNGEROCHEN, UNGETREU oder UNGELEHRIG.

Die Konsequenz, die sich daraus ziehen ließe, wäre etwa, die Eintragungen der mit UN präfigierten Wörter auf solche Fälle zu beschränken, bei denen sich die Bedeutung und die syntaktische Funktion nicht *allein* aufgrund einer allgemeinen Regel erschließen ließe (die etwa lauten könnte: UN-X = (noch) nicht X). Einträge wie UNMELODISCH, UNKRITISCH, UNKAMERADSCHAFTLICH, UNSCHARF könnten entfallen, Einträge Wie UNSELIG (˜ unglücklich, verhängnisvoll), UNSCHEINBAR (˜ nicht auffallend, unbedeutend) würden wegen ihrer nichtableitbaren Bedeutung weiterhin als Worteintrag im Lexikon verbleiben. Schließlich wäre ein genereller Hinweis zur Negierbarkeit (im Sinne von Chomskys Dreiteilung 'möglich und belegt', 'möglich, aber nicht belegt' und 'unmöglich') angebracht. <S. 29>

Das Konzept, nach dem Lexika wie der 'Wahrig' oder der DUDEN aufgebaut werden, ist orientiert an den möglichen Benutzern. Diese an sich triviale Feststellung zeigt aber, daß bei der Erstellung eines derartigen Lexikons nicht die Frage der Eliminierung redundanter Informationen im Vordergrund steht, sondern eine breite Steuung der sachlichen Informationen angestrebt wird. Zugleich sollen möglichst breite Sprachschichten erfaßt werden; daher werden auch die 'wichtigsten' Fremdwörter und Fachbegriffe aufgenommen. Derartige Lexika stellen – zumindest gilt das für die deutschsprachigen Wörterbücher – eher ein Sammelsurium aller

erschienenen Lexika dar, die höchstens auf veraltete Formen hin überprüft oder umneue' Wörter ergänzt worden sind 29/l, als daß sie in ihrem Aufbau und Umfang den linguistischen Ansprüchen, wie sie beispiels= weise die TG formuliert, gerecht werden.

Es ist dagegen bezeichnend, daß gerade in der Linguistischen Datenverarbeitung – aus der technischen Begrenztheit heraus – nach Wegen zu einem Lexikon gesucht worden ist, das möglichst wenig redundante Informationen enthält. Hier werden also auch von der Performanz-Seite der Sprache prinzipiell die gleichen Forderungen gestellt, wie sie unter dem Kompetenz-Aspekt in der TG theoretisch formuliert worden sind. Vereinfacht ausgedrückt lautet diese Forderung:

(6) Als Lexikon-Einträge sind nur solche zuzulassen, die nicht aufgrund von Redundanz- und Wortbildungsregeln durch Komposition anderer lexikoneinträge erzeugt oder aus einem Lexikoneintrag in Verbindung mit einem Affix abgeleitet und dabei in ihrer Funktion erschlossen werden können.

Um den möglichen Aufbau eines Lexikons zur maschinellen Sprachanalyse zu veranschaulichen, orientieren wir uns an dem Saarbrücker Verfahren zur automatischen Erkennung syntaktischer Strukturen beliebiger deutscher Sätze. Aus <S. 30> praktischen Erwägungen wurde bisher 30/l dabei von Wortformen als der morphologischen Komponente des Lexikons ('Stichwort') ausgegangen. Es hat sich gezeigt, daß eine derartige Konzeption nicht befriedigt, da der Aufwand zur Erweiterung des Lexikons beträchtlich ist. Es kommt hinzu, daß das maschinelle Wortformenbuch – gesehen im Verhältnis zu einem die entsprechenden Formen erfassenden Stammwörterbuch – zu umfangreich wird, da die deutsche Sprache sehr flexionsreich ist, Ein günstigeres Konzept scheint daher die Verwendung eines solchen Stammwörterbuchs zu sein, wobei die morphologische Komponente diejenige Graphemfolge darstellt, die bei allen Flexionsformen übereinstimmt, d.h. in der Regel den Grundformen der Wörter abzüglich der Flexionsendungen entspricht (beim Verb etwa dem Infinitiv minus (E)N, beim Adjektiv der unflektierten Form, beim Substantiv dem Nominativ Singular); dieses Konzept wird bei dem in Saarbrücken eben angelaufenen Projekt 'Automatische Lemmatisierung' verwendet. 30/2 Bei flexionsreichen Sprachen hat dieses Stammwörterbuch gegenüber einem Vollformenwörterbuch (Wortformenbuch) den Vorteil, daß die benötigte Speicherkapazität – auch heute stehen kaum Computer zur Verfügung, die ein beliebig großes Lexikon im Kernspeicher halten können – beträchtlich reduziert und die Wörterbuchkodierung erleichtert werden können.

Während bei dieser Wörterbuchstruktur die Flexionsendungen anhand eines noch verhältnismäßig einfachen Erkennungsalgorithmus überprüft und die damit verknüpften Informationen auf diesem Weg leicht der Textwortform zugeordnet werden können, belastet ein weiteres Wörterbuchkonzept dieses Regelsystem in weit stärkerem Maße, wobei allerdings das <S. 31> Lexikon noch weiter entlastet wird. Im Idealfall enthält die morphologische Komponente alle möglichen freien Morpheme, auf die die Wortformen des Textes (also auch die Komposita und Ableitungen) anhand eines komplexen Erkennungsalgorithmus, der die entsprechaiden morphologischen Redundanzregeln enthält und dabei auch die gebundenen Morpheme wie Suffixe und Präfixe sowie die Fugenzeichen einbeziehen muß, reduziert werden können. 31/1 Schwierigkeiten ergeben sich hier in bezug auf die Auflösung morphologischer Mehrdeutigkeiten (HÜHNER-EI, ABT-EI), die nicht unbedingt auch semantisch realisiert sein müssen. 31/2 Es wären in diesem Fall eine

große Zahl von speziellen Restriktionsangaben nötig, um falsche Auflösungen, d.h. Informationszuordnungen, auszuschließen. So kann im Hinblick auf generelle syntaktische Homographien die Anwendung einer Redundanzregel durchaus von Vorteil sein, z.B. in bezug auf die generelle syntaktische Mehrdeutigkeit *Adjektiv im Positiv/Adverb im Komparativ* (SCHÖNER, FEINER ...). Es wäre also nur ein Merkmal nötig, das die Möglichkeit der Komparativbildung selbst beinhaltet. Bei einigen Adjektiven wird der Komparativ jedoch unregelmäßig gebildet (GUT – BESSER; HOCH – HÖHER); hier müssen dann entsprechende morphologische Restriktionsangaben eine entsprechende Beschränkung anzeigen, 31/3 während bei Adjektiven wie BIOLOGISCH, KERNPHYSIKALISCH ein entsprechender Hinweis zu geben wäre, daß hier überhaupt keine Steigerung möglich ist.

Darüberhinaus kennt eine natürliche Sprache wie das Deutsche eine große Zahl syntaktischer und semantischer Mehrdeutigkeiten, die nicht an Stämme oder Morpheme, sondern an eine Wortform gebunden sind. Sie lassen sich im allgemeinen nicht systematisieren. Etwa ist das Wort <S. 32> EHE nur in dieser Flexionsform des Substantivs (zufällig übereinstimmend mit der Grundform) syntaktisch mehrdeutig (Substantiv / Konjunktion, falls die Großschreibung unberücksichtigt bleibt – wie am Satzanfang), die semantisch mehrdeutige Grundform TOR ist in der Flexionsform TORE eindeutig klassifizierbar, die Wortform BILLIGE (Adjektiv / Verb) ist in der unflektierten Form BILLIG oder bei BILLIGT eindeutig der Wortklasse nach bestimmbar. Um diese Mehrdeutigkeiten zu erkennen, darf beispielsweise bei Verwendung eines Stammwörterbuchs nach einer erfolgten eindeutigen Zuordnung der Erkennungsprozeß nicht abgeschlossen werden, sondern die Zuordnungsversuche müssen (evtl. mithilfe eines besonderen Mehrdeutigkeitsmerkmals) noch fortgesetzt werden. Es wäre auch möglich, in diesen Fällen die Wortform ins Lexikon aufzunehmen und die entsprechende Mehrdeutigkeit dabei unmittelbar zu vermerken. Dadurch ließe sich der Suchprozeß verkürzen, die Lexikonerstellung würde aber erschwert, da bei jedem Eintrag alle Flexionsformen auf potentielle Mehrdeutigkeiten hin verglichen werden müßten, eine Arbeit, die der Computer allerdings ebenfalls durchführen könnte. Letztlich hängt die Entscheidung für diese oder jene Version davon ab, inwieweit Wörterbuchsuche (Zeit) und Wörterbuchu~ifang zueinander ins Verhältnis gesetzt werden.

Zum Problem der Mehrdeutigkeiten und zur Frage der Behandlung unregelmäßiger Flexionsformen tritt nbch ein weiterer Umstand, der einen konsequenten (gleichartigen) Aufbau eines Lexikons nicht sinnvoll erscheinen läßt, vor allem die Möglichkeit eines reinen Morphemlexikons ad absurdum führt: Während Wörter wie HAUSDACH oder KÜCHENFENSTER aufgrund ihrer freien Morphemkomponenten HAUS + DACH bzw. KÜCHE + FENSTER semantisch leicht erklärbar sein mögen 32/1, läßt sich kaum noch eine semantisch zutreffende Relation <S. 33> für AUTOSTRASSE oder BÜRGERMEISTER anhand der Komponenten aufstellen. Der Grundsatz der generativen Grammatik, daß Merkmale (also auch die phonologischen) dann im Lexikon aufgeführt sein sollen, wenn sie idiosynkratisch, also nicht in allgemeinen Regeln erfaßbar sind, zwingt dazu, ein maschinelles Lexikon aufzubauen, das die drei Konzepte eines (a) rein an freien Morphemen, eines (b) an Stämmen und eines (c) an Wortformen (oder Teilstämmen) orientierten Wörterbuchs je nach Notwendigkeit und Möglichkeit verbindet.

2.2.1.2 Die Merkmal-Komponente des Lexikons

Die syntaktisch-semantische Merkmalklassifikation ist – wie verschiedentlich zu zeigen versucht wurde – an dem Axialyseverfahren zu orientieren. Zuvor müssen wir uns jedoch mit der Frage auseinandersetzen, ob die Wörterbuchstruktur von der Vorgehensweise des Systems, also von Richtung und Ziel des Algorithmus beeinflußt ist. Mit anderen Worten: Unterscheidet sich ein Lexikon für ein Analyseverfahren von einem Lexikon für ein Syntheseverfahren? 33/1

Wenn man davon ausgeht, daß das zu analysierende Textmaterial nur 'grammatische' Sätze enthält – eine in anwendungsorientierten Parsern häufige und wie ich meine dabei durchaus legitime Annahme – ist man zu folgern versucht, auf eine Reihe von Restriktionen, besonders semantischer Art, verzichten zu können. 33/2 Entweder tritt der Satz DIE KATZE FRISST DEN ELEFANTEN im Text nicht auf – oder er ist eben grammatisch. Es scheint zunächst, als könnte man hiermit die allgemeine Ansicht der Transformationsgrammatiker widerlegen, daß zwischen einer Expansionsgrammatik (Synthese) und einer Reduktionegrammätik (Analyse) kein Unterschied bestehe: Synthetisch betrachtet ist ein Satz wie dieser nicht erlaubt oder im Vergleich <S. 34> zu dem Satz DIE KATZE FRISST EINEN VOGEL weniger grammatisch: bestimmte Restriktionen müssen seine Erzeugung also verhindern. Bei der Analyse könnte man etwa voraussetzen, daß er nicht vorkommt; tritt er dennoch auf, wird er ebenso analysiert wie der grammatische Satz DIE KATZE FRISST EINEN VOGEL. Das Problem ist hier mit dem Wort ebenso charakterisiert: Es steht außer Zweifel, daß in dem 'ungrammatischen' Satzbeispiel keine bei Katzen übliche Nahrungszuführung ausgedrückt werden soll. Diese Unterscheidung zwischen normaler und unüblicher Nahrungsaufnahme würde bei der angegebenen Analyse nicht deutlich werden, so daß die zugrundegelegte Grammatik weniger adäquat ist als eine diese Restriktion berücksichtigende Analyse. Zudem wäre der Satz DIE KATZE FRISST EINE MAUS oder der invertierte Satz EINE MAUS FRISST DIE KATZE ohne entsprechende semantische Restriktionen nicht aufzulösen oder mehrdeutig.

Wir können in diesem Zusammenhang nicht viel mehr feststellen, als daß es nötig ist, zu einem detaillierten semantischen Merkmal- und Analysesystem zu gelangen, um diese deutliche Schwäche der bisherigen Reduktionsalgorithmen zu beheben.

Günstiger und bedeutend leichter zu systematisieren weil weniger komplex – sind die Merkmale zur syntaktischen Komponente der Lexikoneintragung. Zunächst werden in der Regel den Lexikoneintragungen grammatische Kategorien zugeordnet. Je nach Art der Untergliederung läßt sich der Wortschatz einer Sprache in etwa sechs bis zwanzig sog. Wortklassen gruppieren. Die hier genannten Zahlen sollen nur die unterschiedlichen Darstellungsmöglichkeiten symbolisieren, ansonsten spielen sie keine Rolle. Beispielsweise kann man funktional argumentieren und keine allgemeine Wortklasse 'Verb' ansetzen, sondern gleich gliedern in Infinitiv, Finites Verb, Partizip, ... Man unterscheidet im allgemeinen Wortklassen mit vorwiegend funktionalem und sehr geringen semantischen Inhalt <S. 35> wie Konjunktionen, Präpositionen, Pronomina und solche mit spezifischem semantischem Inhalt wie Substantive, Adjektive und Verben. 35/1 Diesen Kategorien lassen sich weitere (Selektionsbeschränkende) Subklassifizierungen zufügen, wobei diese Subkategorisierungsmerkmale nicht unbedingt in streng hierarchischer Relation zu den grammatischen Kategorien stehen müssen. 35/2 Daneben sind Merkmale aufzunehmen, die sich auf notwendige oder mögliche Eigenschaften des Kontexts beziehen (Rektionshinweise, Perfektbildung mit SEIN

oder HABEN...); schließlich sind Angaben zum Erkennen und Anwenden von Transformationen nötig (wie 'Passivbildung möglich'). Sofern die Angaben zu Genus, Kasus und Numerus oder Tempus und Modus nicht anhand morphologischer Redundanzregeln erzeugt werden können, sind sie (z.B. bei notwendigen Wortformeneintragungen) im Lexikoneintrag zu berücksichtigen.

2.2.2 Der Erkennungsalgorithmus

Wir behandeln in der Folge nur den linguistischen Aspekt der Zuordnung von Lexikoneintragungen zu bestimmten in Texten auftretenden Wortformen und lassen die technischen Möglichkeiten und Voraussetzungen der Realisierung (alphabetische Ordnung, Hochfrequenzwörterbuch, Anlage verschiedener, etwa anhand der Wortklassen aufgegliederter Wörterbücher, verschiedene Verfahren eines Dictionary-Look-Up) außer Betracht. 35/3 Die Struktur des Mischwörterbuchs (Wortformen, Stämme oder Morpheme können als <S. 36> morphologische Komponente auftreten) läßt im Prinzip zwei Verfahren zu:

(a) Sind alle mehrdeutigen Wortformen anhand einer ein zigen Komponente – also der Wortform selbst – im Lexikon aufzufinden, kann man eine generelle Regel einführen, die den Charakter einer Redundanzregel hat: Innerhalb der Lexikoneintragungen herrscht eine streng hierarchische Ordnung. Beim Vergleich eines Textwortes mit dem Lexikoneintrag hat der Wortformeneintrag Vorrang vor dem Stammformeneintrag, und dieser muß wieder vor dem Morphemeintrag berücksichtigt werden. Auf diese Weise lassen sich Restriktionsangaben ersparen, die falsche Ersetzungen verhindern sollen. Dazu einige Beispiele: Tritt in einem Text die Wortform SORGE auf, so wäre eine Zuordnung bei dem Wortstamm SORG (Verb SORGEN) möglich. Die Wortform SORGE ist jedoch syntaktisch mehrdeutig: sie steht zugleich für das abstrakte Substantiv SORGE. Ähnlich ist es mit der Wortform VERLEGEN, die unter dem Verbstamm VERbE&:eingeordnet würde, wobei die Information verlorenginge, daß diese Wortform auch Adverb/ Adjektiv sein kann, (ER BLICKTE/IST VERLEGEN), falls keine derartigen Prioritäten geschaffen wären. Hätte die Morphemanalyse Vorrang vor der Stammformenbestimmung, müßte man etwa Restriktionen vorsehen, die verhinderten, daß Wortformen mit spezifischer Bedeutung wie BÜRGERMEISTER (die im Lexikon als Komposita mit entsprechenden Merkmalen eingetragen sein müßten, dem gleichen Zuordnungsprozeß unterworden würden wie die evtl. semantisch auch aus den Komponenten erschließbaren Wörter SCHNEIDER/MEISTER oder TISCHLER/ MEISTER, bei denen semantisch-lexikalische Redundanzregeln angewendet werden könnten. <S. 37> Die Verwendung einer Hierarchieregel führt also zur Einsparung von Restriktionsspezifikationen .37/1

(b) Eine weitere Möglichkeit wäre – wie schon in Abschnitt 2.2.1.1 angedeutet – die Mehrdeutigkeit nicht als solche im Lexikon (unter dem Wortformeneintrag, falls nur Teile eines Lemmas betroffen wären, unter einem Lemma bei semantisch mehrdeutigen Lemmata) zu verzeichnen, sondern die Erkennung dem Zuordnungsalgorithmus zu überlassen. Die Voraussetzung dazu ist ein entsprechender Aufbau der morphologischen Matrix; alle zuordnungsfähigen Stämme – u.U. also auch gleichgeschriebene Stämme – müßten dabei als morphologischer Eintrag vorhanden sein. 37/2 Diesmal darf die Zuordnung nach eihem positiven Ergebnis nicht abgebrochen werden, sondern die Suche muß solange weitergehen, bis aufgrund der morphologischen Komponente, d.h. der Grapheme, keine positiven Zuordnungen mehr zu erwarten sind

> (Schlimmstenfalls könnte erst nach dem Beginn eines neuen Anfangsbuchstabens des Alphabets abgebrochen werden). Während nach Vorschlag (a) etwa bei dem Wort EHE nur ein morphologischer Eintrag erfolgen müßte (da dort u.a. auch alle Homographiemerkmale verzeichnet sind), müßte nach Vorschlag (b) das Wort EHE zweimal im Lexikon stehen – einmal mit den Angaben 'Konjunktion ... ', zum andern mit den Substantivmerkmalen. Das Erkennungsprogramm (b) stellt also zwei (oder mehrere) gültige Lösungen fest und kla3sifiziert das Wort daher als (entsprechenden) Homographen. <S. 38>

Wenn auch nach Vorschlag (b) die Anzahl der aufzunehmenden Wortformen beträchtlich reduziert würde, käme man dennoch wegen der graphematischen Irregularitäten des Sprachsystems nicht völlig ohne Wortformen- oder Teilstammeinträge aus; Wortformen wie BESSER (Steigerung von GUT), MEHR (Steigerung von VIEL) sind idiosynkratisch und müssen in jedem Falle als morphologische Komponente aufgenommen werden. Die Folge davon ist, daß ansonsten regelmäßige Formen wie VIEL u.U. mit Restriktionsmerkmalen versehen werden müssen, falle lexikalische Redundanzregeln verwendet werden (Beispiel: VIEL + ER – 'nur Positiv', während die Redundanzregel lauten würde 'unflektiertes Adjektiv/Adverb + ER – Positiv (attributiv) oder Komparativ (unflektiert), wie sie etwa bei Einträgen wie SCHÖN, LIEB... angewendet werden könnte.

Zunächst ist es also die Funktion des Erkennungsalgorithmus, einer Textwortform mithilfe einer passenden morphologischen Komponente des Lexikons alle Merkmale anzugliedern, die sich über den Lexikoneintrag unmittelbar oder mithilfe der über Flexionsendungs-, Umlautoder Ableitungs- und Kompositionsmerkmalen arbeitenden Redundanzregeln mittelbar erschließen lassen. Zu den Redundanzregeln sind vor allem solche zu rechnen, die einer Textwortform aufgrund von Deklinationsmerkmalen oder Konjugationsangaben der Lexikoneintragung über den Vergleich der Endungen bestimmte Eigenschaften in Bezug auf Kasus und Numerus bei Nomina bzw. Person, Numerus, Tempus und Modus bei Verben zufügen. Zu den Redundanzregeln können darüberhinaus auch solche gehören, die gebundene Morpheme zur Kategorienunterscheidung oder -ermittlung benutzen. Man denke an Ableitungseilben wie -ISCH, -LING oder UNG, die – wo es möglich ist, die 'neue Bedeutung' aus den Komponentenmerkmalen zu erschließen (so z.B. bei KIND-ISCH, NEID-ISCH; FRECH-LING, ROH-LING; BEOBACHT -UNG, VERHAFTUNG) – durchaus anhand <S. 39> von semantischen Redundanzregeln klassifiziert werden könnten.

Der Erkennungsalgorithmus muß dabei Restriktionen unterworfen werden, die im Lexikoneintrag spezifiziert sein könnten durch (a) allgemeine Merkmale des Lexikoneintrags, d.h. solche, die auch zu anderen Zwecken – etwa der Sprachanalyse – herangezogen werden (etwa die Kategorie 'Substantiv' bei Eintragungen wie KIND und NEID) oder (b) durch Merkmale, die nur für diese Redundanzregeln des Algorithmus von Bedeutung sind (etwa der Hinweis, daß bei der Verwendung einer morphologischen Komponente als Spezifikans eines Kompositums stets ein 'S' in der Wortnaht stehen muß. 39/1

Betrachtet man den Aufbau der beiden Lexikonkomponenten nun noch einmal aus dem Blickwinkel des Erkennungsalgorithmus, so zeigen sich die gleichen Konsequenzen, wie sie schon am Beispiel der Starken Verben (Kap. 2.1.3) erläutert worden sind. Falls bei der morphologischen Zuordnung Wortbildungsregeln in Form von Ableitungsund Kompositionsregeln angewendet werden, läßt sich der

morphologische Eintrag auf Kernmorpheme 39/2 beschränken. Dieses Vorgehen ist solange sinnvoll, wie sich das ParsingSystem auf die Erkennung syntaktischer oder morphologischer Gegebenheiten beschränkt. Sobald jedoch feinere (semantische) Subkategorisierungen erfolgen und bei der Analyse berücksichtigt werden sollen, muß konsequenterweise eine idiosynkratische, also auf einen Einzelfall zutreffende 'Regel' erstellt oder aber das bisher 'reine' <S. 45> Morphemwörterbuch erweitert werden um solche Einträge, deren Subkategorisierung sich nicht mehr anhand von Bildungsregeln erschließen läßt bzw. bei denen eine Erkennungsregel zu falschen Resultaten führen könnte. ESELSBRÜCKE, BAHNHOF, BÜRGERMEISTER und MILCHSTRASSE müßten also als Stichwort im Lexikon stehen, während Wörter wie UNSICHER, SCHNEIDERMEISTER, SPEZIALFALL oder ZUGSCHAFFNER evtl. anhand der beteiligten Kernmorpheme über eine Regel auch in ihrer Bedeutungsrelation zu erschließen wären.

3. Zur Konzeption eines Basislexikons

'Welche lange mühselige Arbeit hat dieses Werk mir auferlegt, welchem Gram und Kummer, welchen Kränkungen und Verletzungen mich ausgesetzt, welche Opfer von mir gefordert!' – so äußerte sich Graff in der Vorrede zu seinem 'Althochdeutschen Sprachschatz' 40/1, dessen Vollendung nach über zwanzigjähriger Arbeit er nicht mehr erleben durfte. Dies ist ein noch relativ bescheidenes Beispiel dafür, daß die Herstellung eines umfassenden Wörterbuchs stets mit langwieriger und oft mühsamer Arbeit verbunden gewesen ist. Hundert Jahre hat es gedauert, bis das 1854 von Jacob Grimm begonnene 'Deutsche Wörterbuch' schließlich abgeschlossen war, 40/2 die Neuauflage, 1965 in erster Lieferung zum ersten Band begonnen, ist bisher (Stand 1971) bis zur dritten Lieferung gediehen – das letzte erfaßte Wort ist ABLASZ. 40/3 An derartigen Lexika zu arbeiten erfordert mühselige Kleinarbeit, und daran hat sich auch wenig geändert, als moderne technische Verfahren – vor allem zur Belegerstellung und -sortierung – entwickelt wurden.Auch heute würde eine umfassende Bearbeitung großer Textmengen zum Zwecke der Wörterbucherstellung noch <S. 41> Generationen von Lexikographen beschäftigen können, wollte man sich nicht auf eine möglichst repräsentative Textauswahl beschränken. 41/1 Ein vollständiges Wörterbuch ist ebenso unvorstellbar wie eine vollständige Grammatik. Je größer die Beschränkungen sind, denen man sich bei der Textauswahl unterwirft, desto weniger adäquat – bezogen auf den idealen Sprecher/Hörer – ist auch das Lexikon.

3.1 Begrenztheit der lexikoneintragungen

Fragt man nach den Ursachen der Begrenztheit eines realen Wörterbuchs, das etwa an den Zielsetzungen der TG orientiert ist, so lassen sich zwei Gesichtspunkte unterscheiden: Einmal ist das Lexikon ein Teil der sprachlichen Kompetenz; es beschreibt also die Verwendungsmöglichkeiten, die ein Sprecher/Hörer hat. Zugleich gibt es einen bestimmten Zustand des Sprachsystems wieder, der vielleicht am Tage X um n Uhr gegolten hat. So betrachtet, ist das Lexikon unbeweglich. Doch die Welt und die menschlichen Vorstellungen, die das Lexikon in gewisser Weise widerspiegelt, sind veränderlich. Eine Forderung an das Wörterbuch ist also, daß es sich mit den Vorstellungen ändert, daß es modifizierbar, dynamisch ist. Auf der anderen Seite bestimmt der Verwendungszweck die Struktur des Lexikons. Heute werden mehr und mehr Fachwörterbücher (Spezialwörterbücher) entwickelt, also Lexika, die auch die Stichwörter – nicht nur die zugeordneten Informationen – verwendungsorientiert auswählen; der Umfang des Lexikons ist also von der sprachlichen <S. 41> Performanz her bestimmt. Dies ist besonders deutlich bei sog.

'veralteten' Wörtern: Wenn zu erwarten ist, daß ein Lexikoneintrag nicht mehr nachgeschlagen oder verwendet werden wird, entfernt man ihn wieder aus dem Lexikon. 41/1 Zu diesem Komplex sind auch die primär sprachlich orientierten Lexika zu rechnen, die den Grundwortschatz zu einer Sprache (auch den in einem bestimmten Alter oder für eine bestimmte Sprachschicht notwendigen Wortschatz) vermitteln oder darstellen wollen. 41/2

3.1.1 Deskriptive und explanative Adäquatheit

Wir haben festgestellt, daß sich kein Lexikon erstellen läßt, das die Kompetenz des idealen Sprecher/Hörersbeschreibt. Eine ideale deskriptive Adäquatheit ist also auch nicht für das Lexikon für eine maschinelle Sprachanalyse zu erreichen; wohl aber ein gewisser Grad an Adäquatheit, der sich mit der Kompetenz eines realen Sprecher/Hörers vergleichen ließe.

Das computerorientierte Lexikon wird in der Regel dadurch aufgebaut, daß die augenblickliche sprachliche Kompetenz eines oder mehrerer Menschen, möglicherweise auch die Kompetenz bereits vorhandener Wörterbücher, im Bereich der zu erstellenden Informationen (also nicht unbedingt das ganze enzyklopädische Wissen des Informanten betreffend) in computerzugängliche Form gebracht wird. Das Lexikon kann erweitert werden, wenn sich die Kompetenz eines Bearbeiters entsprechend erweitert hat. Bestenfalls stellt das maschinelle Lexikon also die Vereinigung der bekannten Lexikoneintragungen (und evtl. den Wissens) der menschlibhen Bearbeiter dar. <S. 43>

Wenn eine Grammatik dem Anspruch auf eine möglichst vollständige Beschreibung der sprachlichen Zusammenhänge genügen soll (oder ein Parser in der Lage sein soll, beliebige Sätze einer Sprache adäquat zu analysieren), müssen notwendig Wege beschritten werden, die aus dem Dilemma der statischen Beschränktheit von Regelsystem und Lexikon herausführen. Chomsky 43/1 hat am Beispiel des Erlernens einer Sprache durch ein Kind versucht, diesem Phänomen in der Grammatiktheorie gerecht zu werden. Er nimmt an, daß ein Kind eine intuitive, also angeborene, Kenntnis von sog. 'linguistischen Universalien' besitzt (Beispiel für Universalien, die allen Sprachen gemeinsam sind, ist etwa das Vorkommen von Eigennamen) und daß es außerdem über eine Strategie verfügt, mit deren Hilfe es anhand der sprachlichen Daten, die ihm zugänglich werden, die Grammatik (s)einer natürlichen Sprache aus der im Grunde bereits angelegten universalen Grammatik 'auswählt'

Ob man der Argumentation Chomskys bis dahin folgen will, daß ein Kind bereits eine allgemeine Grammatiktheorie besitzt, oder ob man annimmt, daß eine allgemeinere Prädisposition des Menschen (Lernfähigkeit) für die Grammatikerstellung verwendet wird: Was in dieser 'Black Box' letztendlich geschieht, ist weniger wichtig als das Ergebnis: Es ist unbestreitbar, daß ein Kind bereits Möglichkeiten besitzt} sein sprachliches Regelsystem und seinen Wortschatz aufzubauen, zu erweitern und ggf, zu korrigieren. 43/2 Wortschatz und Regelsystem sind also sowohl statisch (und damit deskriptiv faßbar: zu einem bestimmten Zeitpunkt liegt ein entsprechend abrufbares System vor) als auch dynamisch (und damit explanativ) faßbar: geeignete Strategien erlauben eine Modifikation des Systeme, normalerweise unter Verwendung primärer sprachlicher Daten, wobei u.U. die (physikalische) Umwelt als Korrektiv und Informationsvermittler einbezogen wird. <S. 44>

Trotz dieser Erkenntnisse beschränkte sich die TG im wesentlichen auf den deskriptiven Bereich der Grammatik, wenn auch mit dem ernstzunehmenden Argument, durch die Ermittlung sprachlicher Universalien zunächst die Basis für die Erforschung der Spracherlernung erstellen zu wollen. Das Problem eines dynamischen *Regelsystems* hat m.E.in der Tat keine vorrangige Bedeutung, da sich die sprachlichen Regeln offensichtlich über einen längeren Zeitraum nur unwesentlich verändern und damit Zeit genug lassen, sie in einem deskriptiven System zu ergänzen; sie sind also weitgehend statisch, so daß zumindest im morphologisch-syntaktischen Regelbereich des Sprachsystems eine nahezu vollständige Beschreibung möglich erscheint. 44/1

Anders dagegen verhält es sich mit dem Problem des Umfangs eines *Lexikons*. Denkt man an die numerisch kaum abschätzbare Vielfalt der Eigennamen (Familiennamen, Ortsnamen,...), an die fast ebenso zahllosen Fachtermini und an die große Zahl seltener, d.h. ungebräuchlicher Wörter, so wird die Unmöglichkeit deutlich, ein umfassendes Gesamtlexikon zu erstellen. Selbst wenn dies gelänge – etwa auf der Basis eines Morphemwörterbuchs – so bleiben noch die Fremdwörter und solche Neologismen, die sich aus bisher nicht verwendeten sprachlich möglichen-Graphemkombinationen zusammensetzen. 44/2 Diese Möglichkeit, beliebige 'aussprechbare' Graphemfolgen zu verwenden (zumindest in Form einer Substantivierung oder eines Eigennamens (Der Hase KRUXLIBUX, das Wort SIDRAGO); Beispiele sind vor allem die Kunstnamen bei Industrieprodukten wie FAKT (Waschmittel), NIVEA (Creme)), wirft die wichtige Frage auf, ob ein Lexikon überhaupt feste Grenzen haben darf. Jedenfalls handelt es sich um Einheiten, die nicht in ein festes (Basis-)Lexikon aufzunehmen sind, sondern in einem flexiblen Kurzzeitlexikon <S. 45> verzeichnet sein könnten – ein in der maschinellen Sprachanalyse durchaus praktizierbares Verfahren. 45/1

3.1.2 Der Umfang des maschinellen Lexikons aus technisch-ökonomischer Sicht

Bei der Frage nach dem Aufbau eines maschinellen Lexikons kommen zu diesen allgemeineren, prizipiell linguistischen Argumenten noch weitere hinzu: Der Aufwand der Herstellung eines Lexikons hängt zunächst ab von der Komplexität der einzufügenden Informationen. So kann z.B. ein Lexkon zur syntaktischen Analyse schneller erstellt werden als ein Lexikon, das semantische Spezifikationen enthält, die es etwa erlauben sollen, Polysemien aufzulösen oder syntaktische Mehrdeutigkeiten zu entscheiden. Darüberhinaus ist der Herstellungsaufwand direkt proportional der Anzahl der Lexikoneinträge. Ein Lexikon mit 10.000 Einträgen benötigt also – den gleichen Schwierigkeitsgrad vorausgesetzt – die Hälfte der Zeit, die für 20.000 Eintragungen erforderlich ist. Diese an sich triviale Feststellung 45/2 hat durchaus ihre praktischen Konsequenzen, relativiert man die Dimensionen etwas: Wenn ein an den häufigeren Wörtern orientiertes Lexikon mit 10.000 Eintragungen 98 % aller fortlaufenden Wörter eines beliebigen Textes erfassen könnte und eine Steigerung auf 99,5 % nur durch die Aufnahme von weiteren 50.000 (!) Eintragungen möglich wäre, 45/3 ist zu fragen, ob sich der Aufwand an Arbeitszeit dafür noch lohnt oder ob sich nicht durch geeignete Auswahl der Eintragungen (etwa nach Häufigkeitsgesichtspunkten) und zusätzliche Prozeduren eine brauchbarere, praktikablere Lösung finden ließe. <S. 46>

Ähnliche Argumente gelten für die Wörterbuchsuche, also die Zuordnung von Textwortformen zu einem Lexikoneintrag. Die einfachste und schnellste Zuordnung wäre dann gegeben, wenn man eine Wortform als direkte Adresse (evtl. versehen mit einem multiplikativen Faktor) für die entsprechende Wörterbuchinformation

betrachten könnte. Selbst wenn ein derart überdimensionaler Wörterbuchspeicher zur Verfügung stünde – bei der möglichen Länge der Wortformen eines Textes und aufgrund der Kombinationsmöglichkeiten der Alphabetzeichen (selbst unter Optimierungsgesichtspunkten, d.h. unter Berücksichtigung sprachlich nicht möglicher Zeichenkombinationen wie AFGRTWSTO oder WWERRDFF) bereits eine Utopie – würden in diesem Speicher große Informationslücken entstehen, da eben nicht jede sprachliche Lücke im Code berücksichtigt werden kann. Eine Wörterbuchsuche ist in der Regel also keine einfache Adressenrechnung, sondern besteht – evtl. in Kombination mit einer Adressenrechnung, die etwa das Anfangsgraphem als Beginnadresse verwertet – aus Vergleichoperationen. Je mehr Eintragungen in einem maschinellen Lexikon stehen, desto mehr Vergleiche (wenn auch nicht direkt proportional) sind durchzuführen. Je größer ein computerzugängliches Lexikon also ist – ganz abgesehen von der *einmal* zu leistenden Mehrarbeit bei der Informationserstellung – um so zeitraubender wird die Wörterbuchsuche, weitgehend unabhängig davon, welche Zuordnungsstrategien verwendet werden.

Darüberhinaus sind die maschinellen Wörterbücher im allgemeinen – eben wegen ihres beträchtlichen Umfangs – nicht im Speicher mit der optimalen 'Zugriffszeit' des Computers, also dem Kernspeicher, untergebracht, sondern befinden sich auf den peripheren Speichern der Rechenanlage, also auf Magnetband oder Magnetplatte; auch dies bedingt wiederum längere Such- und Verarbeitungszeiten. Der Umfang des Lexikons bestimmt also neben der evtl. Einschränkung auf <S. 47> die dann noch möglichen (längsameren) physikalischen Speichermöglichkeiten wesentlich die benötigte Rechenzeit (=Kosten) und die Verweilzeit (=die Zeit, in der das Programm im Speicher ist: bei Time-Sharing werden zwischenzeitlich, also vor allem dann, wenn ein Programm auf einen peripheren Speicher zugreift, andere Programme verarbeitet, die evtl. erst dann wieder unterbrochen werden, wenn sie ihrerseits ein derartiges Gerät ansprechen). Es geht dabei natürlich weniger um die Frage, ob man 3.000, 8.000 oder 11.ooo Lexikoneinträge zulassen soll, als um das Problem, lo.ooo oder 500.000 Eintragungen vorzunehmen. Zu derartigen, wenn nicht noch weitaus größeren Dimensionen dürfte man gelangen, wenn man alle Fachausdrücke, alle möglichen Fremdwörter, alle Ortsund Flurnamen oder etwa die in allen Telefonbüchern der Welt verzeichneten Eigennamen notieren und einbeziehen wollte.

Zu dem linguistischen Problem der Begrenztheit eines deskriptiven Lexikons kommen also zeit- und kostenorientierte Argumente für eine Einschränkung des Wörterbuchumfangs.

3.2 Vorschläge für eine Mindestausstattung des Lexikons

Es kann wohl kaum mehr bezweifelt werden, daß für eine relevante automatische Informationserschließung sprachlicher Daten ein maschinelles Lexikon erforderlich ist, sofern man sich nicht bei der Analyse der Daten auf statistische Untersuchungen explizit eingegebener Merkmale (etwa Wortschatz- oder Graphemstatistiken, präkodierte Wortklassen) beschränken will. Eine syntaktische oder semantische Analyse, schließlich eine automatische Übersetzung, setzen in jedem Falle ein Lexikon voraus, da auf Informationen aufgebaut werden muß, die sich aus der morphologischen Oberfläche allein ganz selten erschliessen lassen. <S. 48>

Nach den Darlegungen in Abschnitt 3.1 ergibt sich aufgrund der geforderten Begrenzung des Lexikons die Frage nach einer optimierten Lösung im Hinblick auf

seinen Umfang. Es soll hier allerdings nicht der Versuch gemacht werden, Richtlinien für einen zahlenmäßig optimalen Umfang zu ermitteln, d.h. irgendwelche äußeren Grenzen zu ziehen. Dazu müßten umfangreiche Textstatistiken geführt werden, die zum einen aufgrund der vielschichtigen Auswertungsprozeduren den Rahmen dieser Arbeit sprengen würden, zum anderen mit dem wohl niemals befriedigend lösbaren Problem eines für eine Sprache 'repräsentativen' Textmaterials stehen und fallen würden. Es kann hier auch nicht darum gehen, Untersuchungen zu führen über einen hinreichenden Zuordnungsanteil von Textwörtern zu Lexikoneinträgen, d.h. also etwa darüber, ob bereits eine Texterfassung von 95 % als hinreichend für eine Optimiertheit angesehen werden kann, oder ob dies erst bei 99,9 % der Fall sein kann: Dies sind Probleme, die erst empirisch, also bei einer konkreten Anwendung gelöst werden können und von der individuellen Bewertung des Benutzers abhängen. Ein Übersetzungsbüro ist vielleicht erst mit den 99,9 % zufrieden, da dadurch vielleicht der Mitarbeiterstab erst reduziert werden kann; ein wissenschaftliches Institut begnügt sich möglicherweise mit den 95 %, da es sowieso mit einer maschinellen 'Rohübersetzung' vorlieb nehmen muß, bei der die Mängel eines unvollständigen Wörterbuchs weniger ins Gewicht fallen als die Mängel des Parsing-Systems.

In dieser Arbeit sollen methodische Wege vorgeschlagen werden, deren Anwendung eine sinnvolle Optimierung des Lexikonumfangs erlaubt. Wenn von der Dynamik des Sprachschatzes gesprochen worden ist, so trifft dies doch nicht für alle Elemente in gleichem Umfang zu. Es wurde bereits festgestellt, daß sich der Regelschatz einer Sprache offensichtlich nicht so schnell ändert, sich also nicht von einem Tag auf den anderen, sondern vielleicht auf <S. 49> Jahre oder Jahrzehnte gesehen wandelt. Ein gleiches gilt für die Elemente einiger Wortklassen der Sprache, die gelegentlich – da sie vorwiegend syntaktische *Funktionen* ausüben und keine (oder geringe) eigenständigen Bedeutungsträger darstellen – als Funktionswörter oder Leerwörter bezeichnet werden. Zu ihnen werden die Konjunktionen (UND, DASS,...), die Post- und Präpositionen, die 'reinen' Adverbien (Zeit-, Ort-,...), die Partikel und Interjektionen (JA, OH,...) und die verschiedenartigen Pronomina einschließlich der Artikel gerechnet. Die absolute Anzahl der Elemente dieser Wortklassen ist verhältnismäßig gering; es gibt im Deutschen vielleicht 250 – 300 Präpositionen 49/1 und einige Dutzend Konjunktionen; auch die Anzahl der Pronomina ist klein und – was noch wichtiger ist – vollständig aufzählbar, also endlich. In den meisten herkömmlichen Grammatiken findet sich eine vollständige (oder doch fast vollständige) Liste dieser Wörter. Sieht man einmal von den mehrwortigen Ausdrücken ab, die in ähnlicher Funktion verwendet werden können 49/2 (z.B. DEN GANZEN TAG, AUS DIESEM GRUND) – auch ihre Zahl ist im Grunde überschaubar und abgrenzbar -, so sind diese Wortklassen in Funktion und Bedeutung als wenig flexibel, auf eine kurze Sprachepoche bezogen sogar als statisch anzusehen. Ihre vollständige Aufnahme in das Lexikon scheint aus diesen Gründen bereits sinnvoll zu sein;hinzu kommt, daß Sie in der Regel zu den am häufigsten gebrauchten Wörtern einer Sprache gehören.

Zu den verbleibenden Grundwortklassen (-Kategorien) der Sprache wären also – ohne hier über Grenzfälle oder die Wortklasseneinteilung rechten zu wollen – die Elemente folgender syntaktischer Kategorien zu rechnen: (Finite und infinite) Verben, Substantive, Eigennamen, Adjektive <S. 50> und (Adjektiv-) Adverbien. Ähnlich wie bei den Adverbien geschehen, scheint es noch sinnvoll, eine Subkategorie der Verben, nämlich die der Hilfe- und Modalverben, auszunehmen, da sie ebenfalls eher eine statische als dynamische Menge darstellt, läßt man einige Übergangsfälle (BRAUCHEN, PFLEGEN,...) außer acht. Von einigen, in Kap. 3.2.2

noch zu behandelnden Fällen abgesehen, wäre damit der invariable Teil des optimierten Lexikons bereits umgrenzt.

Einen zweiten festen Bestandteil bilden die Ausnahmelisten; es handelt sich hier um Elemente der übrigen Wortkategorien, die bei der Anwendung morphologischer Redundanzregeln zu mehrdeutigen (sprachlich irrelevanten) oder falschen Ergebnissen führten oder den Regelteil in zu starkem Maß belasten würden. Dafür einige Beispiele, zunächst eingeschränkt auf die syntaktische Merkmalerkennung:

So kann eine Regel vorgesehen sein, die auf der Ableitungssilbe -UNG (und der entsprechenden Flexionsendung -UNGEN) basiert und zu einer Kategorisierung 'Substantiv femininum' führen würde. Eine solche Regel führte jedoch bei Wörtern wie JUNGEN (Flexionsform des Adjektivs JUNG), auch bei auftretenden (evtl. defektiven) Partizipien wie (GE,ER-)RUNGEN, (GE,ZER)SPRUNGEN, (GE-,...)ZWUNGEN, (BE-,...)SUNGEN oder (GE-,...)WRUNGEN zu falschen, in Fällen wie LUNGEN, ZUNGEN zu zwar ad hoc richtigen, aber trotzdem unbefriedigenden Ergebnissen. Ähnliches gilt für Regeln, bei denen andere Suffixe zu ihnen entsprechenden Kategorisierungen führen; so müßten bei dem Suffix -HEIT etwa die Ausnahmen GESCHEIT und auch SCHEIT verzeichnet werden, bei -IG (Ergebnis : Adjektiv) wären Wörter wie ESSIG, REISIG, (evtl.:) DANZIG, KÖNIG, ZEISIG, HONIG, PFENNIG, MENNIG, ZWEIG, KÄFIG, TEIG, STEIG, FEIG (!) in der Ausnahmeliste zu führen. Das Auflisten weniger Ausnahmen erlaubt hier – im syntaktischen Bereich – die <S, 51> Informationserschließung einer Großzahl von Wörtern anhand *einfacher* morphologischer Regeln, ohne das Lexikon übermäßig zu belasten. 51/1 Sofern zu diesen Funktionalen Informationen weitere Angaben, evtl. zur semantischen Subkategorisierung, erschlossen und verarbeitet werden sollen – sie müssen in der Regel ebenfalls dem Lexikoneintrag zugeordnet sein – muß die Ausnahmeliste um diejenigen Sonderfälle erweitert werden, bei denen eine morphologische Redundanzregel zu Fehlern in der semantischen Zuordnung führen würde. Hier fände das bereits in 2.1.4 angedeutete Problem seine Konsequenzen; BÜRGERMEISTER müßte evtl. (wollte man nicht ein zweites Lexem MEISTER ansetzen) in der Ausnahmeliste verzeichnet, sein, um eine Regel

> X-MEISTER – 'durch Ablegung einer Prüfung als Meister des X-Berufs anerkannt'

zu erlauben. Dabei könnten also die lexikoneinträge TISCHLERMEISTER, SCHNEIDERMEISTER, SCHNEINERMEISTER erspart werden.

Der Umfang der Ausnahmelisten ist also abhängig von Art und Umfang der morphologischen Redundanzregeln. Quantitativ besonders ertragreich ist das Listen der Ausnahmen bei einer Sprachanalyse, die nur bis zur syntaktischen Strukturerkennung vordringen will. Hier genügen wenige Ausnahmen, um über eine Regel eine Vielzahl weiterer Wörter syntaktisch präzuklassifizieren. Sobald eine feinere, auf die Bedeutungsstrukturen und -Relationen vordringende Analyse erstrebt wird, wachsen diese Ausnahmelisten sehr schnell an; bei einer maschinellen Übersetzung scheinen sie – vielleicht abgesehen von der Auflösung einer Reihe von Komposita und einigen Ableitungssuffixen – weiter an Bedeutung zu verlieren, da eine Regel <S. 52> nur noch vereinzelt angewendet werden kann; gänzlich bedeutungslos wird dieses Verfahren aber auch dann nicht sein, wenn die Zahl der notwendigen Lexikoneinträge – denn die Ausnahmeliste ist nichts anderes als ein Teil des Lexikons – auch stark ansteigt.

Prinzipiell ist das Basislexikon mit der Erstellung der Ausnahmelisten abgeschlossen. Dennoch wird man aus ökonomischen Gründen eine weitere Stufe ansetzen müssen, die der Entlastung des Regelteiles dient; Da bei jedem nicht im Lexikon unmittelbar, d.h. über die Graphemfolge, verzeichneten Wort die gesamte Kategorisierung durch das Regelsystem geleistet und das Ergebnis dieser Kategorisierung spätestens nach einem Analyseprozeß wieder 'vergessen' wird, muß diese 'Arbeit' bei jedem erneuten Auftreten 52/1 des Wortes vom Computer aufs neue durchgeführt werden. Treten diese Wörter häufig auf, so schlägt dies in einer Erhöhung des Rechenaufwands evtl. deutlich zu Buche. Eine Aufnahme solcher hochfrequenter Wörter (Morpheme, Stämme, Wortformen) -unabhängig von ihrer systematischen Einordnungsmöglichkeit – scheint daher sinnvoll zu sein.

Diese dritte im Zusammenhang mit einer Optimierung des Lexikonumfangs aufgestellte Kategorie von Lexikoneinträgen ist damit aber bereits geprägt von individuellen, anwendungsorientierten Gesichtspunkten. 'Anwendungsorientiert' nicht im Sinne des verwendeten Grammatiksystems oder der erstrebten Adäquatheit der Grammatik, sondern im Hinblick auf den bei der Analyse zu erwartenden Sprachbereich. 52/2 Da solche Begriffe wie 'Sprachbereich' oder 'Häufigkeit' in bezug auf ihre Grenzen nicht wohldefiniert sind, ist man auf empirische Analyseergebnisse angewiesen. Sofern <S, 53> diese nicht schon festliegen, ist eine Kontrolle der Lexikoneinträge mithilfe von maschinell erstellbaren Statistiken zur Zugriffhäufigkeit eines Lexikoneintrags während der Arbeit an dem Material ein brauchbarer Weg, die Lexikoneinträge zu modifizieren, d.h. solche Wörter (der dritten Lexikonkategorie) auszulagern, die unter einem – empirisch als günstig ermittelten – Häufigkeitswert liegen, andere dem Erkennungsalgorithmus zu entziehen und ins Lexikon zu überführen.

3.2.1 Strategien zur Klassifikation unbekannter Wörter

Aus der gegebenen Begrenztheit der Lexikoneintragungen folgt notwendig die Frage nach der Behandlung v6n Textwortformen, die nicht mithilfe der vorhandenen Lexikoneinträge – seien es nun Wortformen, Stämme oder Kernmorpheme – erkannt und/oder mittels morphologischer Redundanzregeln klassifiziert werden können (Es sei hier abgesehen von dem Problem einer unzureichenden oder fehlerhaften Merkmalklassifikation). Durchaus sinnvoll – vor allem bei einer automatischen Sprachübersetzung – scheint der Aufbau eines (interaktiven) Kommunikationssystems Mensch-Maschine zu sein; in allen Fällen, in denen keine Informationszuordnung mithilfe des Wörterbuchs möglich ist, wendet sich der Computer an seinen menschlichen Partner und erhält von diesem die entsprechenden Informationen (evtl. erfolgt anschließend auch noch eine Erweiterung des Lexikons um den entsprechenden Eintrag). Eine derartige Lösung bietet sich gerade heute an, in einem Stadium der Computertechnik, in dem der Aufbau von Dialogsystemen (mithilfe von Datensichtgeräten) bereits praktizierbar ist und ökonomisch gerechtfertigt erscheint. Das Ergebnis wäre also eine fast vollautomatische Sprachverarbeitung (-Analyse, -Übersetzung); der Mensch bildet dabei gleichsam ein peripheres Ersatzlexikon für den Computer. <S. 54>

So praktikabel diese Methode auch erscheint 54/l, sie ist aus der Sicht des Linguisten noch unbefriedigend. Die zweite Möglichkeit, die sich anbietet, ist die der Integration des nicht im Wörterbuch oder Regelsystem 'enthaltenen' Wortes in das Klassifikationsverfahren der maschinellen Analyse. Diese Vorgehensweise baut im Grunde auf der Vorstellung auf, daß ein unbekanntes Wort aufgrund der übrigen Textsituation, also des (bekannten) Kontexts, klassifizierbar sein kann.

Wenn wir beispielsweise den Satz lesen

HERR DAKAPOPULOS HAT SEINEN ONKEL BESUCHT.

so nehmen wir an, daß DAKAPOPULOS ein Name ist (Wir wissen dabei etwa, daß HERR als Anredeform vor Namen steht usw.). In dem Satz

ER BEWEGTE SICH DEGAGIERT.

kann man – ohne die semantische Struktur des Satzes genau zu erkennen – das letzte Wort zumindest als adverbial gebrauchtes Partizip bestimmen. Den beiden Beispielen ist gemein, daß aufgrund der Stellung und des bekannten syntaktischen oder semantischen Kontexts – evtl. auch aufgrund besonderer morphologischer Merkmale des Wortes selbst – eine gewisse Klassifikationsstufe eines unbekannten Wortes erreicht wird. 54/2 Es ist nun denkbar, daß ein unbekanntes Wort in einem Text mehrere Male vorkommt <S. 55> (vielleicht ist der Themenkreis nicht allgemeinsprachlich), daß dabei ein anderer Kontext die Merkmalklassifikation noch erhärtet oder spezifiziert. Diesen gesamten Vorgang könnte man als eine einfache Form des Lernens, nämlich des Lernens durch Erfahrung, 55/1 apostrophieren. Wenn es nun gelänge, für einen Computer einen derartigen Regelmechanismus aufzubauen, müßte es zumindest bis zu einem gewissen Grad der Klassifizierung möglich sein, die Kluft zwischen einem endlichen Lexikon und den – wenn man so will – unendlichen Wortbildungsmöglichkeiten zu überbrücken.

3.2.2 Die Funktion des Kontexts

Auf der Basis des beschriebenen Lexikons läßt sich in Verbindung mit den morphologischen Erkennungsalgorithmen ein Großteil des Wortschatzes einer Sprache bewältigen, d.h. eindeutig mit den für die weitere Bearbeitung erforderlichen Informationen ausstatten. Ein Teil dieser Zuordnungen führt jedoch aufgrund der 'natürlichen' Mehrdeutigkeiten in der Sprache auch zu (lexikalisch) mehreren Lösungsmöglichkeiten. Zunächst einige Beispiele im syntaktischen Bereich der Wortklassen: EINIGEN kann nicht nur Indefinitpronomen sein, sondern auch finites und infinites Verb; für SEIN (Possessiv-Pronomen/Infinitiv), BRACH (Verbzusatz/finites Verb) gilt ähnliches; vernachlässigt man gar das Merkmal der Groß- und Kleinschreibung des Anfangsbuchstabens (dies gilt in jedem Fall für den Satzanfang), werden diese Mehrdeutigkeiten noch vermehrt. Als Beispiele für semantische Mehrdeutigkeiten seien die allbekannten Wörter DAME (im Schachspiel, Dame-Spiel, 'Frau') und SCHLOSS (auf dem Berg, an der Tür...) angeführt. Eine Klassifizierung dieser Mehrdeutigkeiten ist bereits mehrfach versucht worden 55/2; wir können hier darauf verzichten. <S. 56>

Zur Auflösung dieser Mehrdeutigkeiten benutzt der Mensch – beschränken wir uns einmal auf die Schriftsprache – in erster Linie den Kontext. Dazu ein Beispiel:

PETER HAT DIE DAME GESCHLAGEN.

Dieser Satz ist zugleich syntaktisch und semantisch mehrdeutig: PETER / DAME können Subjekt / Objekt sein; die semantische Mehrdeutigkeit wird deutlich, wenn wir sie durch einen Nebensatz vereindeutigen:

(a) PETER HAT DIE DAME GESCHLAGEN, DA SIE NICHT MEHR VON DEM SPRINGER GEDECKT WURDE.
(b) PETER HAT DIE DAME GESCHLAGEN, OBWOHL DIE POLIZEI IHN GEMAHNT HATTE, NICHT MEHR SO GEWALTTÄTIG ZU SEIN.

In beiden Sätze wird PETER bei diesem Kontext eindeutig als Subjekt bestimmbar. Stünde in Satz (b) statt IHN das Wort SIE, wäre DAME Subjekt und PETER Objekt (wenn die Satzstellung in diesem Zusammenhang auch etwas unüblich wäre). Bei einer maschinellen Analyse wird im Prinzip ebenso verfahren, zur Auflösung dieser Mehrdeutigkeiten also ebenfalls der Kontext herangezögen. Während der Mensch jedoch in der (glücklichen) Lage ist, sein ganzes enzyklopädisches Wissen dabei in die Waagschale zu werfen, muß der Computer mit den relativ bescheidenen Informationen seines Lexikons und seinem verhältnismäßig einfachen Regelsystem dieses sprachliche Phänomen zu bewältigen versuchen. Daß dieses Verfahren durchaus praktikabel sein kann, hat das Saarbrücker Verfahren m.E. deutlich gezeigt, obgleich es noch auf einer bescheidenen syntaktischen Ebene operiert. 56/1

Es stellt sich nun in unserem Zusammenhang die Frage, ob der Kontext nicht auch in ähnlicher Weise für die Klassifikation 'unbekannter' Wörter herangezogen werden kann. Daß dies möglich erscheint, sei an einem kleinen Beispiel verdeutlicht (für das unbekannte Wort steht 'X'): <S. 57>

X HAT SEINEN ONKEL BESUCHT.

Hier kann für X.(mit wenigen Ausnahmen wie den Personalpronomen ER, SIE, ES oder VATER, MUTTER) nur ein Eigenname stehen. Ein anderes (syntaktisches) Beispiel:

DER MANN X AUF DEN BAUM.

Setzt man – wie wir es bei dem Minimallexikon getan haben – voraus, daß alle Funktionswörter und Hilfsverben bekannt sind, wird es sich bei X um ein Verb handeln. Bei dem Satz

DER MANN KLETTERT AUF DEN X.

läßt sich leicht folgern – die ausgeschriebenen Wörter seien bekannt – daß X ein Substantiv (evtl. auch ein SubstantivErsatz, etwa ein (substantiviertes) Adjektiv)sein muß, zumindest wird die Möglichkeit 'Verb' ausgeschlossen. Ich habe hier bewußt einfache Satzbeispiele gewählt und mich auch davor gehütet, weitere Spekulationen über eine mögliche differenzierende Subkategorisierung hier folgen zu lassen. Zugleich wurde von der Annahme ausgegangen, daß der Kontext bereits klassifiziert wurde. Bedenkt man aber den großen Anteil der natürlichen Mehrdeutigkeiten im Deutschen und die mögliche Komplexität der Satzstruktur, schließlich die Möglichkeit, daß mehrere 'unbekannte' Wörter in einem Satz auftreten können, so muß versucht werden, den Anteil der unbekannten Wörter – trotz des begrenzten Lexikonumfangs zu reduzieren oder zumindest die Mehrdeutigkeiten einzuschränken. Ich führe dazu einen Begriff ein, den ich im Gegensatz zur 'natürlichen' Mehrdeutigkeit des Sprachsystems als 'künstliche' Mehrdeutigkeit bezeichnen will.

Die 'unbekannten' Wörter können a priori allen Merkmalklassen, deren Elemente nicht vollständig im Lexikon verzeichnet sind oder die nicht ausschließlich durch entsprechende Regeln erfaßt werden können, angehören. Der Regelfall wird jedoch

sein, daß sie in der sprachlichen Realisierung nur einige spezifische Merkmale aufweisen (Damit soll nicht ausgeschlossen werden, daß sie auch 'natürlich' mehrdeutig <S. 58> sein können). Dennoch sind sie zunächst mit allen noch möglichen Merkmalen auszustatten; dann ist durch die Kontextanalyse – nach ähnlichen Verfahren wie bei der Auflösung natürlicher Mehrdeutigkeiten – eine Reduktion dieser Mehrdeutigkeiten durchzuführen. Veranschaulicht sei dies an dem Beispiel der Wortklassenmerkmale, für das auch die in Kap. 4 dargelegte Strategie entwickelt werden ist. In Kap. 3.2 wurden bereits die 'offenen' Wortklassen Verb, Substantiv, Eigenname, Adjektiv und Adverb angeführt. Die Klasse cbr Eigennamen will ich aus der folgenden Betrachtung ausklammern, da ihre besondere Problematik (auch bekannte `Wortformen können als Eigennamen auftreten : MÜLLER, KLEIN, ZIMMERMANN, POST , KAUFMANN, ENGEL, MAUS, ADLER ..., seltsamerweise nicht BAHNHOFSOBERINSPEKTOR, KRANKENWÄRTER, HÜHNERHUND, EISENBAHN, RAKETE...) besondere Erkennungsstrategien erfordert. 58/1 Es bleiben also vier Hauptklassen, wobei sich das Verb syntax-funktional noch einmal aufgliedern läßt nach finitem Verb, Infinitiv und Partizip (II unflektiert). Wollte man die unbekannten Wörter mit diesen Mehrdeutigkeiten ausstatten, so wären sie in dieser Hinsicht a priori sechsdeutig.

Der eigentlichen syntaktischen Analyse läßt sich aber auch hier ein (quasi-) morphologischer Regelteil vorschalten, in dem aufgrund graphematiseher Merkmale erste Wortklassenreduktionen durchgeführt werden. 58/2 Hierzu ein paar einfache Beispiele, die sich noch erweitern und verfeinern ließen: Geht man wie in Kap. 3.2 gefordert beispielsweise davon aus, daß alle starken Verben (einschließlich der Präfigierungen) im Deutschen über das Lexikon erfaßt werden können, kann für die Kategorie 'Partizip II unflektiert' gelten, daß das letzte Graphem <S. 59> ein 'T' ist. Trifft dies zu, wird zugleich die Möglichkeit des flektierten Adjektivs ausgeschlossen. Endet ein Wort nicht auf N, so kann es kein Infinitiv sein; endet es auf einen Vokal außer 'E', werden Adjektiv, Verb, Partizip ausgeschlossen usw.1 Auf diese Weise läßt sich die 'künstliche' Mehrdeutigkeit in vielen Fällen2 weitgehend reduzieren, so daß die endgültige Vereindeutigung anhand des Kontexts durch den Parser erleichtert wird.

3.2.3 Eine automatische Rückkopplung zur Wörterbucherweiterung oder Informationskorrektur

Wenn man auch damit rechnen muß, daß die unbekannten Wörter *natürliche* Mehrdeutigkeiten darstellen – dies ließe sich durch die Aufnahme der entsprechenden Wörter in das Basislexikon einschränken -, so werden die natürlichen Mehrdeutigkeiten in der Regel eine echte Teilmenge der künstlichen bilden. Im vorigen Abschnitt wurde bereite auf das Problem hingewiesen, daß der engere Kontext – im allgemeinen der Satzzusammenhang – aufgrund seiner komplexen Struktur oder eines zu großen Anteils an (natürlichen oder künstlichen) Mehrdeutigkeiten – nicht immer *sichere* Schlüsse im Hinblick auf die reale Funktion eines unbekannten Wortes zuläßt. Daher ist nach Wegen zu suchen, die diesen Undicherheitsfaktor reduzieren helfen.

Die in der maschinellen Analyse noch auswertbare Information findet sich im Textzusammenhang oder – im weiteren Sinn – im gesamten bisher aufbereiteten Sprachmaterial. Man kann von der Annahme ausgehen, daß ein unbekanntes Wort – vorausg?setzt, es kommt irgendwo in einem Text wiederholt vor einen anderen Kontext im engeren Sinn aufweist, der eine Lösung entweder bestätigt, gegebenenfalls aber auch falsifiziert. Bestätigung bedeutet zugleich 'Verstärkung'

des vorherigen, eine abweichende Lösung <S. 60> schwächt evtl. eine vorheriges Ergebnis in seiner Aussagekraft ab. Ein weiteres Kriterium kann die Wahrscheinlichkeit 60/1 oder 'Sicherheit' einer Reduktion bedeuten. Ein Beispiel dafür ist die Mehrdeutigkeit Relativwort/Demonstrativwort (˜ Artikel). Setzt man voraus, daß auch die Zeichensetzung regelgerecht angewendet worden ist, so gibt es folgende eindeutige, d.h. 'sichere' Regel, die die Möglichkeit 'Relativwort' ausschließt: Steht vor einem Homographen DEM/REL kein Satzzeichen und steht davor keine Präposition oder Kunjunktion, so handelt es sich um ein Demonstrativwort. Hierdurch kann zwar die weitaus schwierigere Lösung Relativwort nicht erfaßt werden, aber mehr als 80 % aller Lösungen für Demonstrativwort lassen sich durch diese 'sichere' kontextsensitive Regel ermitteln. Ähnlich 'sichere' Lösungsregeln gibt es auch für andere Homographen.

Eine Rückkopplung des Ergebnisses einer Mehrdeutigkeitsauflösung mit dem Wörterbuch bedeutet nun, daß das Ergebnis dieser Reduktion wieder festgehalten – und bei einem wiederholten Auftreten u.U. entsprechend verwertet wird. Es ergibt sich beispielsweise die Möglichkeit, 'sichere' Lösungen unmittelbar in das Lexikon (das durchaus ein Kurzzeitlexikon sein kann, also nur für die Analyse eines Textabschnittes aufgebaut wird) zu überführen, falls eine natürliche Mehrdeutigkeit aufgrund der vorhandenen Ausnahmeliste ausgeschlossen erscheint. Wo dies nicht der Fall ist oder wo eine Lösung als unsicher, weniger sicher ... gekennzeichnet wäre (es ließe sich empirisch eine entsprechende Skala aufbauen), ist eine 'Verstärkung' erforderlich, die nur anhand des wiederholten Auftretens des 'unbekannten' Wortes erfolgen kann. (Es wäre übrigens denkbar, daß diese Verstärkung nicht mit dem Auftreten derselben Wortform verknüpft ist, sondern – über eine zwischengeschaltete <S. 61> Lemmatisierung – an die potentielle Grundform (Lemma) gebunden würde. Auf diese weise könnte etwa eine Verstärkung bei einem Adjektiv (Lemma z.B. DEGOUTANT) über auftretende Wortformen (DEGOUTANT, DEGOUTANTES, DEGOUTANTEN) erfolgen. Diese Möglichkeit der 'Lemmatisierung' unbekannter Wörter vor der Übernahme in das (Kurzzeit-) Lexikon bietet eine größere Wahrscheinlichkeit für eine frühzeitige eindeutige Reduktion unbekannter Wörter. Schließlich wäre auch die automatische Löschung einer Information denkbar, etwa dann, wenn sich eine Anzahl sicherer Lösungen (oder auch nur eine) entgegengesetzter Art ergeben haben und es sich bei der zu löschenden um eine einzelne unsichere Reduktion gehandelt hat.

Ich konnte hierzu nur einige theoretische Möglichkeiten aufzeigen, die im Einzelnen noch empirisch-statistisch zu untermauern oder zu variieren wären. Zum Abschluß sei noch auf eine besondere Anwendungsart dieses Rückkopplungsprozesses hingewiesen. Es ist denkbar, daß man das Lexikon nicht um eine weitere – zufällige – Information oder Einträge erweitern möchte, da sonst die Vorteile einer bewußten Begrenzung auf eine Basis verlorengehen könnten. Das Kurzzeitlexikon bietet die Möglichkeit, für die Dauer der Untersuchung eines bestimmten Textes ein weiteres Mittel zur Hand zu haben. Da in vielen Fällen ein größerer Text oder Textausschnitt vor dem Vergleich der Textwortformen mit den Lexikoneinträgen zur Zuordnung der lexikalichen Informationen alphabetisch sortiert wird, um diese Wörterbuchsuche ökonomischer zu gestalten, ist bereits bekannt, wie häufig eine Wortform auftritt. Läßt sie sich nicht über einen Lexikoneintrag zuordnen, so könnte für einen Text ein Zusatzwörterbuch erstellt werden, das vor allem bei wiederholt auftretenden Wörtern – etwa solchen, die allgemeinsprachlich betrachtet selten, textspezifisch aber häufiger sind – einen Verstärkungsprozeß erlauben würde und damit die Möglichkeit gäbe, unsichere <S. 62> Lösungen automatisch zu verifizieren oder aber zu korrigieren. 62/1

Während bisher nahezu ausschließlich über syntaktische Informationen gehandelt wurde, bietet sich für diesen Rückkopplungsprozeß auch eine Anwendungsmöglichkeit für die Lösung semantischer Mehrdeutigkeiten: Er ließe sich in einer Variante auch für Wörter anwenden, die bereits im Lexikon enthalten sind, aber syntaktisch oder semantisch mehrdeutig sind. Im Grunde werden bei diesem Verfahren ja Informationen verwertet, die sich nicht allein aus dem Analyseabschnitt ergeben, der vom Parsing-System gerade bearbeitet wird. Während der Parser in der Regel wohl satzorientiert ist (also eine Satz-für-Satz-Analyse durchgeführt wird), sind diese zusätzlichen Informationen satzübergreifend und textorientiert. Der Reduktionsalgorithmus kann nun so konstruiert werden, daß sich der Computer bei einer 'unsicheren' Lösungsmöglichkeit gleichsam an eine vorherige Lösung 'erinnert' (eben über das Kurzzeitlexikon) und sich dabei u.U. jetzt für eine entsprechende Lösung entscheidet oder sie zumindest als die wahrscheinlichere anbietet. Dazu ein Beispiel: Zu analysieren seien u.a. die Sätze

(a) ER BLICKTE LANGE NACH DEM SCHLOSS HINÜBER, DESSEN MAUERN IM ABENDLICHT SCHINCLERTEN... -
(b) BALD-SAH ER DAS SCHLOSS WIEDER.

Nehmen wir an, daß im Satz (a) eine 'sichere' Auflösung der semantischen Mehrdeutigkeit SCHLOSS (in Richtung auf 'Gebäude') möglich war. Satz (b), für sich allein betrachtet, ist mehrdeutig. In dem übersatzmäßigen Kontext spricht die Wahrscheinlichkeit für die Version 'Gebäude'. Dreht <S. 63> man die ReUnfolge der Sätze um, ließe sich das Verfahren immer noch anwenden: Bei der 'sicheren'Lösung 'erinnert' sich der Computer einer vorher aufgetretenen 'unsicheren' Lösung (oder nicht gelösten Mehrdeutigkeit) und bietet rückwirkend auch dafür diese Reduktionsmöglichkeit mit Priorität an.

3.2.4 Die Behandlung von Eigennamen

Der Rückkopplungsprozeß ließe sich auch für einen Wortschatzbereich verwenden, der bisher aus der Lexikonkonzeption ausgeklammert wurde; den Bereich der Eigennamen. Es ist im Grunde naheliegend, eine Reihe häufig vorkommender Namen gemäß der in Kap. 3.2 vorgeschlagenen Konzeption (3. Lexikonstufe) unmittelbar ins Lexikon aufzunehmen. Ein Teil dieser Namen wird mehrdeutig sein (Substantiv/Name : SCHLOSSER; Adjektiv/Name: GROSS); dazu im Parser entsprechende Reduktionsroutinen vorzusehen, ist jedoch nicht einfach, da besondere die Mehrdeutigkeit Substantiv/Name über kontextsenSitive Merkmale nicht immer sicher aufgelöst werden kann. Dies wird etwa deutlich, wenn man eindeutige Eigennamen durch mehrdeutige ersetzt in Strukturen wie

DER JUNGE GOETHE – DER JUNGE MÜLLER
oder
ER HAT SCHILLER VEREHRT – ER HAT SCHÄFER GESEHEN.

Dennoch bietet sich eine Reihe von Möglichkeiten, in vielen anderen Fällen Reduktionen durchzuführen.1 Etwa ließe sich anhand einer Vornamenliste, einer Berufsbezeichnungsliste, eines Titelverzeichnisses, einer Anredeliste und mithilfe von entsprechenden Wortstellungsregeln eine Reduktion durchführen (GERD MÜLLER, DER MALER CLUESSERATH, DR. GROSS, BISCHOF SCHULZE, HERR ZAUN, auch bei Appositionen: BIRNBAUM, EIN BERÜHMTER MALER

DES 16. <S. 64> JAHRHUNDERTS,...). Diese Regeln wären etwa auch entsprechend bei der Klassifikation unbekannter Wörter zu berücksichtigen. 64/1

Die in diesem Zusammenhang interessante Frage, ob nicht alle Wörter eines Lexikons prinzipiell auch Eigennamen sein können, also eine generelle Regel zur Auflösung der Mehrdeutigkeit X/Eigenname vorgesehen werden müßte, will ich hier nur kurz anschneiden. Es wäre sicherlich eine eigne UntersUchung wert, Regeln zu ermitteln, die erkennen lassen, ob ein Lexikoneintrag als Eigenname verwendet werden kann oder nicht. Diese Möglichkeit einer Trennung läßt sich etwa am Beispiel der Berufsbezeichnungenverdeutlichen: Es scheint einen Zusammenhang zu geben mit dem Alter dines Wortes (Berufs) und der Möglichkeit, als Eigenname verwendet zu werden. 'Alte' Berufe wie MÜLLER, SCHREINER, SCHNEIDER, ZIMMERMANN oder BAUER werden auch als Eigennamen verwendet, 'neuere' Berufe wie INSPEKTOR, ELEKTRIKER, GIPSER nicht; ähnliches gilt für Werkzeuge oder Geräte wie SCHLEGEL, KEIL auf der einen, COMPUTER, BIRNE auf der anderen Seite. Auch neuzeitlichere Wörter wie RUNDFUNK, BAHNHOF treten nicht als Eigennamen auf; vielleicht könnte also die Angabe des Erstbelegs eines Wortes als Restriktion für diese allgemeine Regel herangezogen werden, daß beliebige Wörter einer Sprache zugleich als Eigennamen (auch Ortsnamen sind hinzuzurechnen) auftreten können.

3.3 Zusammenfassung der Ergebnisse

Linguistische und ökonomische Gegebenheiten zwingen dazu, für die maschinelle Sprachbearbeitung ein möglichst kompaktes Basislexikon zu erstellen. Eine Gruppe sprachlicher Kategorien, deren Elemente vor allem formal-syntaktisch verwendet werden (Funktionswörter, Hilfsverben) ist in <S. 65> ihrer Zahl endlich und zugleich weniger umfangreich; es wurde daher vorgeschlagen, sie vollständig ins Lexikon aufzunehmen.

Eine zweite Gruppe, die Menge der 'Ausnahmen' bei der Anwendung morphologischer Redundanzregeln, ist ebenfalls notwendig ins Lexikon zu integrieren, um falsche Lösungen zu verhindern. Schließlich wurde eine dritte (fakultative) Gruppe von Lexikoneintragungen definiert, die es in Abhängigkeit von der spezifischen Verwendungsrichtung des Lexikons – textbezogen – gestattet, durch Aufnahme von häufig auftretenden Wörtern ins Lexikon Rechenzeit einzusparen, die durch die (wiederholte) Anwendung komplexer Redundanzregeln entstehen kann.

Für die Gruppe der nicht eindeutig über Lexikon- und Redundanzregeln erfaßbaren Textwortformen wurde ein Verfahren postuliert, das – ähnlich der Auflösung 'natürlicher' Mehrdeutigkeiten – über eine maschinelle Analyse auf kontextebene eine Klassifizierung der 'künstlich' mehrdeutigen (d.h. in der sprachlichen Realität vielleicht eindeutigen) Wörter versucht. Hier können erste Reduktionen möglicherweise anhand eines quasimorphologischen Regelapparates durchgeführt werden. Schließlich wurde für die Aufnahme dieser 'unbekannten' Wörter ein textorientiertes Kurzzeitlexikon postuliert, das es gestattet, unsichere Lösungen bei häufigerem Auftreten einer Textform (evtl. – falls eine lemmatisierung möglich erscheint – eines Lemmas) – auch im Nachhinein zu verstärken oder umgekehrt auch zu falsifizieren und zu korrigieren. Es ist denkbar, dieses Lexikon auch für andere Zwecke zu verwenden: etwa zur Kopplung neugewonnener Informationen mit dem Batslexikon; d.h. einige Wörter werden aufgrund ihrer Häufigkeit nicht nur automatisch klassifiziert, sondern in das Basislexikon überführt. Schließlich kann

das Kurzzeitlexikon im Zusammenhang mit der Auflösung syntaktischer und semantischer Mehrdeutigkeiten (als übersatzmäßiges Korrektiv) ebenso verwendet werden wie bei der Klassifikation <S. 66> von Eigennamen, sofern mehr als ein entsprechender Textbeleg auftritt.

4 Ein Modell zur automatischen Klassifikation lexikalisch nicht identifizierbarer Wörter

Während in den bisherigen Abshnitten eher theoretisch argumentiert wurde – dies gilt auch für die Konzepte eines Basislexikons und eines Kurzzeitlexikons – und damit der Spekulation mehr Raum gegeben war, wird dieses Kapitel auf die Beschreibung eines konkreten Modells, also nicht auf etwas Machbares, sondern etwas Gemachtes ausgerichtet sein. Ziel dieses Modells war es, die Realisierbarkeit der theoretischen Argumente nachzuweisen; andererseits hat die Arbeit an dem praktischen Verfahren auch auf die theoretischen Formulierungen eingewirkt, d.h. es bestand eine Wechselwirkung zwischen diesen beiden Teilen der Arbeit.

Daß dieses Modell – gemessen an den theoretischen Forderungen – in vielen Fällen unzulänglich ist, wird nicht bestritten. Dies liegt z.T. darin begründet, daß das Analyseverfahren und das Lexikon nicht nur bereits konzipiert, sondern auch praktisch arbeitsfähig waren, ehe mit der Konstruktion dieses Modells begonnen wurde. Weiter sind die Unzulänglichkeiten darin begründet, daß der technische Aufwand, der zum konkreten Nachweis einer theoretischen Forderung im einzelnen hätte geleitet werden müssen, nicht immer in einem angemessenen Verhältnis zur Beweiskraft des Nachweises stand. Dazu ein Beispiel: Es hätte langwieriger sorgfältiger Prüfungen verschiedenster Lexika bedurft, wenn man etwa hätte sichergehen wollen, daß alle Ausnahmen zu einer morphologischen Redundanzregel auch im Lexikon verzeichnet sein würden. Morphologische Fehlklassifizierungen (insgesamt waren es bei den 62 über Redundanzregeln eindeutig klassifizierten Wörtern des <S. 67> Tests 7, von denen 6 über die in der Analyse vorgesehen Substantivierungsregel 67/1 wieder automatisch 'korrigiert' wurden) wie etwa bei OHRFEIGE (Suffix IG) oder ZISCHEN (Suffix ISCH) zeigen eher die Notwendigkeit an, solche Wörter in das Lexikon aufzunehmen, als daß sie einen Beweis für die Undurchführbarkeit des Systems darstellten. Dieses kleine Beispiel zeigt aber, daß von der Postulierung bis zur Realisierung ein weiter Weg ist: es unterstreicht zugleich den Modellcharakter des realisierten Systems.

4.1 Das Saarbrücker Verfahren zur maschinellen Syntaktischen Analyse 67/2

Ehe das konkrete System der automatischen Erkennung unbekannter Wörter in seinen Einzelheiten und im Zusammenwirken seiner Teile jedoch erläutert werden soll, sind die theoretischen und praktischen Voraussetzungen aufzuzeigen, auf denen es aufbaut. Erklärtes Ziel des Saarbrücker Vorhabens war die syntaktische Analyse *beliebiger* deutscher Sätze. Dennoch wurden sowohl das lexikon als auch das Analyseprogramm nicht von vornherein an diesem Ziel orientiert; Ausgangsbasis war vielmehr ein 'repräsentativer' Querschnitt aus Sachprosa und Zeitungen. 67/3
Während das Lexikon (ein Wortformenlexikon) zunächst strikt an diesen Daten (= Wortformen der zu verarbeitenden Texte) ausgerichtet wurde (ca. 30.000 der Wortformen erfassen alle etwa 200.000 laufenden Textwortformen der beiden Texttypen mit zusammen ca. 11.000 Sätzen; weitere 10.000 wurden über Textvergleiche von in Saarbrücken vorliegenden weiteren <S. 68> Prosatexten automatisch aufgenommen), wurden die natürlichen Mehrdeutigkeiten nicht an den

in diesen Texten auftretenden Klassen orientiert, sondern unter Berücksichtigung der möglichen Klassifizierungen festgehalten, wobei man sich also auf die Kompetenz des KodiererLinguisten stützte. Ähnlich wurde bei der Ermittlung der Analyeeregeln verfahren: hier bot die Vielfalt der Texte genügend Möglichkeiten, Regeln für komplexe Strukturen zu ermitteln, ohne daß allerdings alle in den 11.ooo Sätzen auftretenden Strukturtypen oder -Varianten in der Kürze der zur Verfügung stehenden Zeit hätten grundlegend berücksichtigt werden können; auch hier spielte stets – unabhängig vom Text – die Kompetenz des Linguisten eine Rolle, wobei allerdings die 'Akzeptabilität' mit berücksichtigt wurde, also die Frage, ob eine theoretisch grammatische Struktur aufgrund ihrer Komplexität nicht mehr in der sprachlichen Realität auftreten kann;d.h. solche Strukturen wie

> ER HÄTTE EINEN HUND 'LAUFEN SEHEN GEKONNT ZU HABEN SCHEINEN MÖGEN'

wurden außer Acht gelassen.

Die Analysestrategie war dabei an den durch die Wörterbuchsuche gewonnenen Informationen dergestalt orientiert, daß ein Satz nicht analysiert werden konnte, wenn eine Wortform nicht im Lexikon vorhanden oder zwar im Wortlaut vorhanden war, aber (noch) die notwendigen Informationen fehlten. Im Vordergrund des Saarbrücker Verfahrens stand die Entwicklung des Regelsysteme für eine Analyse, so daß es weniger auf die Vollständigkeit des Lexikons – geschweige denn auf eine Systematisierung der Lexikoneintragungen – ankam als darauf, daß – waren alle Wortformen eines Satzes über ein Lexikon mit Informationen ausgestattet – die Analyseregeln zu 'richtigen' Ergebnissen gelangten. Die Analyse beliebiger Sätze wurde also nicht als ein Wortschatz- und damit Lexikonproblem, sondern als ein Struktur- und damit <S. 69> ein.Regelsystemproblem betrachtet. Fehlten beispielsweise zu einem Satz entsprechende lexikoneinträge, so wurde das Lexikon ergänzt; nach einer erneuten Wörterbuchsuche ließ sich der entsprechende Satz analysieren.

Es würde hier zu weit führen, die einzelnen Analyseschritte zu erläutern; dies ist an anderer Stelle 69/1 schon ausführlich geschehen; nur die für diese Arbeit wichtigsten Stufen werden an entsprechender Stelle behandelt. Festzuhalten ist, daß der Lexikonstruktur in dieser Phase nur der zweite Pl$z zukam. Dies heißt nun aber nicht, daß die Notwendigkeit eines Lemikons für die Analyse geringgeschätzt oder daß dem *Inhalt* der Lexikoninformatiohen im einzelnen wenig Beachtung geschenkt worden wäre; stets wurde, vom Informationsgehalt her gesehen, das Lexikon als Teil des gesamten Analysesysteme betrachtet. 69/2

Je mehr sich das Analysesystem verfestigte, desto deutlicher wurde das Fehlen eines systematische(re)n Lexikons offenbar; vor allem zeigte sich dieser Mangel beim Testen von Satzstrukturen. Das an den 11.000 Corpus-Sätzen orientierte Lexikon enthielt Einträge wie AUTOBAHNEN (es fehlte AUTOBAHN) oder SAGTE (es fehlte SAGTEST); die Zahl der Beispiele ließe sich beliebig vermehren.

Diesem Mangel generell abhelfen zu wollen, hätte eine völlige Umstrukturierung, besser gesagt, eine neue Konzeption des Lexikons und der Lexikonstrategie bedeutet, 69/3 zudem wäre das erstellte, immerhin 40.000 Wortformen umfassende Lexikon nicht oder kaum noch weiter als Basis für das neue Konzept verwendbar gewesen,. Mir schien daher ein anderer Weg gangbarer, der den Vorteil mit sich brachte, daß das Wörterbuch in seiner Struktur nicht verändert zu werden brauchte

und auch weiterhin voll verwendbar war, <S. 70> der aber aus dem Engpaß herausführen sollte, daß Sätze mit 'unbekannten', d.h. lexikalisch nicht präklassifizierbaren Wörtern nicht analysierbar waren. Dieser Weg schisn zudem in kürzerer Zeit realisiert werden zu können, während eine völlige Umorganisation des Lexikons nicht mehr innerhalb des für die Syntaxanalyse bewilligten Forschungszeitraums hätte verwirklichst werden können. 70/1

4.1.1 Zum Aufbau des Saarbrücker Lexikons 70/2

Ehe auf die Möglichkeiten eingegangen werden kann, die Kapazität der Voranalyse (=Wörterbuchsuche) zu erhöhen, soll der bisherige Aufbau des Wortformenbuches in Grundzügen erläutert werden:

Die Wortformen werden in bezug auf ihre mögliche(n) Funktion(en) in einem konkreten Satz klassifiziert. Ein eindeutiges Gruppierungsmerkmal ist dabei die distributionale Äquivalenz, d.h. die (hier formal auf die Syntax einzuschränkende) Austauschbarkeit von Wortformen. Wenn im folgenden daher die (auf den ersten Blick vergröbernden) Einteilungsschemata in einer Übersicht dargestellt werden, so ist stets festzuhalten, daß eine eindeutige Klassifizierung einer Wortform erst durch die Interpretation aller (subkategorisierenden) Informationsmerkmale gegeben ist, die der entsprechenden Wortform durch die Voranalyse mitgegeben werden.

Eine erste Unterscheidung bringt bereits die Gruppierung nach Wortklassen. Ohne auf die Problematik der Wortklassendefinition und der verwendeten Begriffe einzugeben, stelle ich die verwendete Einteilung kurz anhand von Beispielen <S. 71> vor. Für die weitere Untersuchung ist es nützlich, sie außerdem nach den semantischen Merkmalen 'inhaltsschwach' (=Funktionswortklasse) und 'inhaltsschwer' (=Bedeutungswortklasse) zu trennen.

(a) Funktionswortklassen

PER	Personalpronomen	ICH, IHM
FRA	Fragewort	WER, WO,
DEM	Demonstrativwort	DIESER, DA
POS	Possessivpronomen	MEIN
IND	Indefinitpronomen	KEIN, EIN
ITJ	Interjektion	AHA
PRP	Präposition	AN, IN
POP	Postposition	WEGEN (!), HALBER(!)
REL	Relativwort	DER, WELCHER
VZS	Verbzusatz	AN, TEIL (!)
KON	Konjunktion	UND, DASS

(b) Bedeutungswortklassen

VRB	Finites Verb	GIBT
INF	Infinitiv	GEBEN (!)
IZU	Infinitiv mit ZU im Wort	ANZUGEBEN
PTZ	Partizip II unflektiert	GESEHEN
SUB	Substantiv	HAUS
NAM	Name	PETER, LAVATER
ADJ	Adjektiv, Partizip (flektiert)	JUNGE, SINGENDE, GESCHEHENE
ADV	Adverb, Partizip I unfl.	SCHON, HEUTE, SINGEND

Zu den Merkmalen der Wortklasse (oder der Wortklassenmehrdeutigkeit, die durch eine Homogra.phennummer angezeigt wird – etwa bedeutet HO 43 die Mehrdeutigkeit DEM/REL-) kommen bei den Nomina die Kasus-, Numerus- und Genusangaben; beim Verb werden Modus, Tempus und Numerus notiert. Außer durch diese Informationen werden einzelne Wortklassen oder <S. 72> Wortformen verschiedentlich weiter spezifiziert durch sogenannte Sonderangaben. Dies sind z. B. Hinweise auf mögliche Rektionen oder Valenzen; Konjunktionen werden je nach ihrer besonderen Funktion als neben- oder unterordnend gekennzeichnet; beim Partizip II wird angegeben, ob es mit den Hilfsverben SEIN oder HABEN eine Verbalgruppe bilden kann usw. Diese syntaktischen Angaben bringen insgesamt also eine genauere Gruppierung der Wortform im Hinblick auf die Möglichkeit ihrer Position und Abhängigkeit im Satz. Eine semantische Subkategorisierung ist für dieses Lexikon nicht gegeben.

4.1.2 Die Auflösung syntaktischer Mehrdeutigkeiten

Auch hier soll nur das zum Verständnis des folgenden Notwendige dargelegt werden, da das Saarbrücker Verfahren schon wiederholt 1 geschildert worden ist. Wichtig ist, zu wissen, daß die Auflösung der Wortklassenmehrdeutigkeit – jene Mehrdeutigkeit, die auch für die erste, grobe syntaktische Klassifizierung einer Wortform wesentlich erscheint und die im Mittelpunkt des hier zu schildernden Modells stehen wird – als eigner, selbständiger Schritt vor der weiteren Analyse des Satzes stattfindet, die sich anschließende Analyse (Nominale und Verbale Gruppierung, Klassifikation von Subsätzen und schließlich des Satzes) also auf den dort 'gewonnenen' Informationen aufbaut. Als Ergebnis der Homographieprogramme wird für jede ;90rtform stets nur eine Lösung angeboten, d.h. eine der in Kap. 4.1.1 beschriebenen Wortklassen. Nur für zwei Fälle ist im späteren Verlauf der Analyse überhaupt noch eine Korrekturmöglichkeit vorgesehen, nämlich bei der Mehrdeutigkeit Relativwort/Demonstrativwort(=HO 43) und bei der Mehrdeutigkeit Verb/Infinitiv/Substantiv (=HO 2). Als einziges, sehr schwaches und fragwürdiges Kriterium <S. 73> im Hinblick auf die Sicherheit oder Unsicherheit einer Lösung kann der sog. Homographendurchlaufzä.hler (D) gewertet werden; bei jedem homographen Wort wird festgehalten, ob es unabhängig von anderen noch nicht gelösten Homographen (etwa aufgrund der Stellung ...) gelöst werden konnte oderr ob es in Abhängigkeit von einem zu diesem Zeitpunkt ungelösten Homographen reduziert werden mußte. Im letzteren Fall (ohne auf weitere Differenzierungen hier eingehen zu wollen 73/1 wird ein Wert $D \geq 6$ festgehalten.

Die Homographenauflösung ist das Hauptproblem der Syntaxanalyse I. Hier wird ein wesentlicher Teil der gesamten Analyse gleichsam vorweggenommen (bis hin zu als vorläufig angenommenen Subsätzabgrenzungen und nominalen/verbälen Strukturerkennungen), da aufgrund der möglichen Komplexität der Sätze (Einbettungen von nominalen Subgruppen oder von Subsätzen, Diskontinuität der Verbalgruppen...) in vielen Fällen nicht die nähere Umgebung, also einige Wörter oder Satzzeichen links oder rechts des aufzulösenden Homographen, ausreicht, eine Mehrdeutigkeit zu reduzieren. Entsprechend anfällig ist auch das Homographenreduktionsprogramm: je länger (d.h. in der Regel komplexer) ein Satz ist, um so fehleranfälliger istdas Lösungsprogramm. Dies zeigte eindeutig eine entsprechende Untersuchung2. Der Anteil der fehlerfrei gelösten Sätze (fehlerfrei im Hinblick auf die Homographie-Auflösung) liegt bei den (kürzesten) 4-Wort-Sätzen bei 94 %, er sinkt über 77 % bei den (durchschnittlich langen) 16-Wort-Sätzen auf 68 % bei den (längsten) 32-Wort-Sätzen.

4.1.3 Die morphologischen Redundanzregeln

Von dem Problem der Homographenauflösung weitgehend unberührt, dafür um so mehr von der äußeren, vom Stichwort her <S, 74> bedingten Ad-hoc-Struktur des Lexikons betroffen sind die morphologischen Redundanzregeln. Falls keine Zuordnung über das Lexikon möglich ist, werden folgende drei 'morphologischen' Regeln in der unten angezeigten Reihenfolge auf alle diese überbleibenden \ortformen angewendet;

(a) Alle ',Wortformen', die alphabetisch gesehen vor dem ersten tatsächlichen Lexikoneintrag stehen, also kleiner sind als das erste Wort im Lexikon, werden als 'Zahl'-Adjektiv betrachtet.

Hier wird sich des Umstands bedient, daß allen Ziffern vom Computersystem ein niedrigerer Wert zugeordnet wird als den Buchstaben. So steht die Zahl 617 vor dem ersten hypotetischen Lexikoneintrag 'A' (in unserem konkreten Lexikon ist es das Wort A-GRUPPE).

(b) Alle Wortformen, die aus (der Kombination von) Morphemen wie EIN, ZWEI, DREI, ..., NEUN, ZEHN, ELF, ZWOELF, ZWANZIG, DREISSIG, ..., HUNDERT, TAUSEND, ... (zugelassen sind noch UND sowie Suffixe wie STE, TE) werden als Zahl-Adjektive betrachtet.

Ein Problem, das in diesem Zusammenhang besonders deutlich zu zeigen ist, ist das der 'Substantivierung' von Adjektiven. In dem Satz

HUNDERTTAUSEND SAHEN DAS FUSSBALLSPIEL.

ist schwer zu entscheiden, ob hier eine Ellipse (MENSCHEN, ANHÄNGER,...) vorliegt oder ob hier von der Möglichkeit Gebrauch gemacht wurde, eine Zahl als selbständige Nominal-’gruppe' zu verwenden. Das Analyseprogramm wendet nach der Homographenauflösung eine (grobe) Substantivierungsregel für Adjektive an, die etwa folgendermaßen verbal umschrieben werden kann:

> "Folgt auf ein Adjektiv kein weiteres Adjektiv oder kein Substantiv oder kein Komma, wird das Adjektiv als 'substantiviert' angesehen (die Wortklasse ADJ wird in SUB umgewandelt)".

Obgleich eine differenziertere Regel, die den weiteren <S. 75> Kontext (evtl. über die Satzgrenzen hinaus) berücksichtigen würde, natürlich bessere Ergebnisse liefern könnte, lassen sich quantitativ gesehen schon durch diese einfache Regel recht brauchbare Resultate erzielen. Wie schon erwähnt, führte diese Regel auch zur Korrektur morphologisch falsch klassifizierter Wörter; nur wenige gegenteilige Fälle (z.B. ELFTE als SUB in Satz 23 der Testsätze1 75/1) sind allein aufgrund dieser Regel fälschlich klassifiziert worden.

(c) Die nicht nach den Regeln (a) oder (b) klassifizierten Wortformen werden einer Suffixanalyse unterzogen; sie erhalten diejenigen Informationen (wie Wortklassen und Flexionsangaben), die bei Übereinstimmung der Graphemfolgen eines Eintrags der Suffixliste mit den letzten Graphemen der Textwortform bei dem Eintrag des entsprechenden Suffixes verzeichnet sind. 75/2

In der Erkennungsstategie ist die Forderung berücksichtigt, daß die Zuordnung bei der größtmöglichen Übereinstimmung erfolgt. Eine Vollständigkeit der Liste zu erreichen war nicht beabsichtigt; in die Suffixliste integriert sind Graphemfolgen oder Grapheme, die die Vereindeutigung einer Wortform zulassen, ohne eine 'Ableitungssilbe' im streng morphologischen Sinn zu sein (etwa führen X, Z, Q, W oder Y als letztes Graphem zur Wortklasse NAM (Eigenname) 75/3).

Daß dieses Prinzip – Ausnahmen ins Lexikon, Suffixe in die Regelliste – keineswegs unproblematisch ist, zeigen solche Suffixe wie REICH, BAR, HAFT oder MAL, die alle mit der Klassifikationsangabe ADJ oder ADV (je nach Flexionsmerkmal) ausgestattet wurden. Es handelt sich hier zugleich <S. 76> um Graphemfolgen, die freie Morpheme darstellen, die als solche wiederum nicht zur Ableitung, sondern zur Komposition verwendet werden können (KÖNIGREICH, HIMMELREICH, LIDOBAR, TANZBAR, VOLLSTRECKUNGSHAFT, VORBEUGEHAFT, EHRENMAL, DENKMAL, WUNDENMAL, ...); es ist zweifelhaft, daß hierzu eine vollständige 'Ausnahmeliste' erstellt werden kann dergestalt, daß alle möglichen Komposita ins Lexikon aufgenommen würden. Die Großschreibung der Substantive als Unterscheidungsmerkmal – hier ein *praktikabler* Ausweg – ist linguistisch wenig befriedigend, abgesehen davon, daß am Satzanfang keine Unterschiede zu den Adjektiv/Adverben gemacht werden. In verschiedenen Fällen tritt die Mehrdeutigkeit nicht bei allen Flexionsformen auf (ZAHLREICHER, WUNDERBAREN, -BARES...); hier läßt sich also dadurch eine Eindeutigkeit erreichen, daß Restriktionen für die Anwendung der Suffixregel vermerkt und beachtet werden. Grundsätzlich müßte aber bei derartig mehrdeutigen Wortformen auf eine vereindeutigende Suffixanalyse verzichtet werden, um Fehler zu vermeiden; diese Wörter sind also der gleichen Prozedur zu unterziehen, die für die noch verbleibenden 'unbekannten' Wörter vorgesehen ist.

4.2 Das System zur Klassifikation unbekannter Wörter

Ein Teil der Textwörter wird weder durch das auf den aufgenommenen Wortformen basierende Erkennungs- und Zuordnungsprogramm erfaßt noch läßt er sich eindeutig anhand morphologischer Kriterien klassifizieren. Zwar können aufgrund einer quasimorphologischen Untersuchung gewisse Wortklassen in vielen Fällen bereits ausgeklammert werden (denken wir an das soeben aufgeführte Beispiel REICH: die Mehrdeutigkeit könnte hier auf SUB/ADV bzw. bei REICHES auf SUB/ADJ eingeschränkt werden; wenn der letzte Buchstabe ein A ist, kann es sich nicht mehr um ein finites Verb oder einen Infinitiv handeln; wenn er nicht N ist, kann es kein Infinitiv mehr sein ...; dennoch ist eine endgültige oder eindeutige Reduzierung durch die morphologische Analyse allein nicht möglich.<S. 77>

In diesem Falle wird das nicht näher bestimmbare Wort als entsprechend mehrdeutig gekennzeichnet, Nach der Ausstattung mit Informationen zur Deklination und Konjugation und mit solchen Informationen, die alle denkbar möglichen syntaktischen Zuordnungen gewährleisten (Informationsüberfluß), durchläuft diese Wortform die Satzanalyse. Während dieser Analyse wird – vergleichbar der Homographenauflösung versucht, eine im konkreten Satz mögliche ('richtige') Funktion. zu ermitteln. Als Ergebnis wird eine Lösung angeboten, die sogar über die verbale oder nominale GruppierungSmöglichkeit (Numerus, Genus, Kasus usw.) Auskunft geben kann.

Diese Lösungen werden einem Kurzzeitlexikon angeboten. Dabei wird das Analyseergebnis über ein geeignetes Verfahren, das auch Sicherheitsgrade der Lösung einbezieht (hier in Abhängigkeit vom Schwierigkeitsgrad des Satzes, ermittelt anhand der Größe des Homographendurchlaufzählers unter Berücksichtigung der Komplexität des Satzes), in dieses Zwischenlexikon überführt.

4.2.1 Die Struktur des Kurzzeitlexikons

Während bei der Auflösung der natürlichen Mehrdeutigkeit das Ergebnis nach der Satzanalyse wieder vergessen wird – es handelt sich dabei nur um eine Verifizierung einer bereits als möglich festgestellten Funktion – muß das Kurzzeitlexikon (im folgenden mit Lexikon bezeichnet) Möglichkeiten vorsehen, die Ergebnisse der Reduktion künstlicher Mehrdeutigkeiten nicht nur zu speichern, sondern auch mit den evtl. bereits erarbeiteten Informationen in geeigneter Weise zu verknüpfen.Es ist damit zu rechnen, daß das Ergebnis aus verschiedenen Gründen, etwa wegen der Unv6llkommenheit der Lösungsmethode oder auch wegen des Überflusses an Information bei diesen 'unbekannten' Wörtern, die zu syntaktisch erlaubten Zuordnungen und damit formal richtigen Klassifizierungen führen, nicht dem möglichen natürlichen <S. 78> Gebrauch der Wortform entspricht. 78/1 So bietet gerade das bisher in Saarbrücken verwendete Verfahren nicht genügend Anhaltspunkte dafür, ob und bis zu welchem Grad eine Lösung als sicher angesehen werddn kann; solche Merkmale in das bestehende Verfahren zu integrieren, wäre einer völligen Neukonzeption der Homographenanalyse gleichgekommen. Es kommt hinzu, daß deutsche Sätze komplexe Verschachtelungen aufweisen können, daß ein Satz aus einer großen Zahl von Wörtern mit natürlicher Mehrdeutigkeit bestehen oder weitere nicht präklassifizierte Wörter enthclten kann. Man muß also davon ausgehen, daß die durch die Analyse angebotene Lösung falsch sein kann.

Eine einzige Reduktion darf also im allgemeinen nicht zur endgültigen Bestimmung der Wortfunktion führen. Tritt ein Wort dagegen mehrfach in verschiedenen Sätzen (in – wie anzunehmen ist – verschiedener Kontextsituation) auf und wird aufgrund der Analyse jeweils (vielleicht auch mit wenigen als unsicher erkannten Ausnahmen) die gleiche Lösung angeboten, wird gefolgert, daß sie einer natürlichen Wortfunktion_ entspricht. Ein Wort kann jedoch auch eine natürliche Mehrdeutigkeit besitzen. Also muß berücksichtigt werden, daß mehrere verschiedene Lösungen richtig sein können.

Alle diese grundlegenden Faktoren sind bei der Überführung der Informationen in das Lexikon zu berücksichtigen. Das Problem der generellen Unsicherheit eines Ergebnisses zwingt bei unserem Verfahren dazu, nicht bereits eine, sondern mehrere (gleiche) Lösungen als 'richtig' zu akzeptieren, d.h. von einem statistisch zu ermittelnden Schwellenwert (Mindestlösungszahl) auszugehen. Dieser Schwellenwert ist prinzipiell abhängig vom Schwierigkeitsgrad des Satzes und der Art der Auflösung (Wortklasse). Seine experimentell zu bestimmende Größe ist beeinflußt von der absoluten Wahrscheinlichkeit des Auftretens einer Information <S. 79> (Wortklasse) im Deutschen und der relativen, d.h. vom Lexikon her bestimmbaren wortschatzabhängigen Vorkommenswahrscheinlichkeit. Sind beispielsweise alle im Deutschen möglichen Verben bereits über das Lexikon präklassifizierbar, ist die Wahrscheinlichkeit, daß die unbekannte Wortform ein Verb ist, =0. Pür den durchgeführten Test wurde diesem Phänomen nicht näher Rechnung getragen, sondern für alle Wortklassen ein gleicher Schwellenwert (vier sichere Lösungen) vorgesehen.

Ich nenne die Phase, in der die Lösungen noch nicht die erforderliche Mindestzahl erreicht haben, die Unsicherheits- oder Lernphase. In dieser Phase werden bei der Analyse eines Satzes nicht die evtl. vorhandenen Ergebnisse vorhergehender Satzanalysen verwertet, sondern das betreffende Wort wird weiterhin als 'unbekannt' betrachtet, mit den allgemein zu gewinnenden künstlichen Mehrdeutigkeitsangaben ausgestattet und entsprechend zu bestimmen versucht.

Wird der Schwellenwert für eine Grundfunktionsangabe, also eine Wortklasse, an einer Stelle überschritten, befindet sich das Wort für diese Lösungsmöglichkeit in der Zwischen- oder Übergangsphase. Während als sicher angenommen wird, daß eine natürliche Wortfunktion festgestellt wurde, ist zu überprüfen, ob andere Funktionen noch möglich sind. Liegen zu anderen syntaktischen Klassifikationen nur 'unsichere' Lösungen vor, so werden diese gelöscht, d.h. vergessen. Auch in der Übergangsphase werden – falls ein unbekanntes Wort eines Textes zugeordnet werden könnte – nicht die Ergebnisse der vorherigen Analysen verwendet oder mitverwendet, sondern wie in der Lernphase die morphologisch möglichen Mehrdeutigkeiten angenommen,

Die Übergangsphase besitzt ihrerseits wieder Schwellenwerte (hier ebenfalls vier sichere Lösungen), die für die entsprechende Wortklassenfunktion neu angesetzt werden, sobald sie in diese Phase eintritt. Wird der betreffende SchweI1enwert erreicht (d.h.: haben sich für eine Wortklasse bereits *acht* sichere Lösungen ergeben), tritt das Wort für diese <S. 80> Wortklasse in die Sicherheits- oder Kannphase über. Alle bisher nicht belegten Lösungsmöglichkeiten werden verboten, 80/1 desgleichen alle Lösungsmöglichkeiten, die nur als unsicher gekennzeichnet waren. Die evtl. vorhandenen weiteren sicheren Lösungen sind von diesem Prozeß nicht beeinflußt. Die Kannphase erlaubt bei einer erneuten Zuordnungsmöglichkeit die Übernahme der ermittelten Informationen der endgültig klassifizierten Wortklassen und die Einschränkung der aufgrund der quasimorphologischen Analyse noch möglichen Deutigkeiten auf die jetzt noch im Lexikon zugelassenen Wortklassen. Bei weiteren Satzanalysen werden diese automatisch endgültig ermittelten Funktionsangaben gleichermaßen behandelt wie die menschlich voranalysierten, dem Wortformenlexikon zugeordneten Wörter.

Wie in Abschnitt 3.2 näher begründet, beschränkt sich dieses Verfahren auf bestimmte Wortklassen. Außer der Reduktion unbekannter Wortformen auf diese Wortklassen (ADV ADJ SUB NAM 80/2 PTZ VRB INF IZU) wird versucht, noch folgende Zusatzangaben durch das Verfahren zu ermitteln 80/3: Genus-, Numerus- und Kasusangaben zu den Substantiven und Eigennamen, die Verwendung des PTZ II in Verbindung mit SEIN oder HABEN (IST... GEKOMMEN, HAT... GESEHEN) und die direkte verbale Rektion. Die Adjektiv-Deklinationsangaben sowie die Konjugationsangaben beim Verb werden mithilfe der Endung morphologisch erkannt und im Lexikon typisiert. Eine Reduktion der mehrdeutigen Konjugationsangaben beim Verb ist nicht möglich, da dieses Problem im Zusammenhang mit der Analyse noch nicht bearbeitet wird. Auf die Ermittlung weiterer Angaben (wie Substantiv oder Verb in Verbindung mit einem Infinitivsatz) muß aus ähnlichen Gründen gleichfalls verzichtet werden. <S, 81>

Vor der Betrachtung weiterer Einzelheiten des Verfahrens sei die Lexikonstruktur mit den darin enthaltenen Informationen in einer kurzen Übersicht dargestellt. Die folgende Tabelle zeigt zugleich den technischen Aufbau des Lexikons und die Form der Speicherung der Analyseergebnisse. Insgesamt werden je Wortform 21 Speicherzellen zu je 27 Bits benötigt. In den ersten 7 Zellen wird der Wortlaut der

Wortform aufbewahrt (je Zelle 4 Buchstaben, maximal also 28 Buchstaben je Wort). Bei allen Wortklassenzellen ist eine 'Sperre' vorgesehen, die angibt, daß die hier evtl. abgelagerten Informationen nicht mehr gültig (d.h. vergessen) sind.

DER AUFBAU DES KURZZEITLEXIKONS

Zelle	*Inhalt*	*Bemerkung*
0 – 6	*Wortlaut*	*drei linke Bits = 1 (Schreibmaschinencode X 1)*
7	ADJ 81/1	Zellenaufbau-für-alle Wortklassen:
8	ADV	Bit 26 (Sperre): Wort kann nicht dieser
9	VRB 81/2	Wortklasse angehören
10	INF 81/3	Bit 21 Ursache der Sperre: nicht belegte WK
11	PTZ	Bit 20 Ursache der Sperre: Graphemfolge
12	SUB	Bits 0-4 Zähler Lernphase unsichere Lösung
13	NAM	„ 5-9 „ „ sichere Lösung
		„ 10-14 „ Übergangsph. sichere L.
		15-19 „ „ unsichere Lösung
14	SUB I D	Deklinationsangaben SUB Lernphase,
15	SUB II D	„ SUB Übergangsphase
16	NAM I D	„ NAM Lernphase, sicher
17	NAM II D	„ NAM Übergangsphase
18	AKK	Rektionsangaben 81/4 (Akkusativ bzw. Dativ):
19	DAT 0-3	Zähler IP PTZ 4-7 Zähler ÜP PTZ
	8-11	„ LP VRB 12-15 „ ÜP VRB
	16-19	“ LP INF 20 -23 " ÜP INF <S 82>
20	SEIN/HABEN	Partizip mit SEIN oder HABEN gebildet 82/5
		0 – 4 Zähler LP SEIN 5 – 9 Zähler ÜP SEIN
		10 -14 " LP HABEN 15-19 " ÜP HABEN
21	leer	

Das größte praktische Problem bildete die konkrete Definition der 'Sicherheit' bzw. 'Unsicherheit' einer Lösung sowie die Relation zwischen beiden Bewertungen. Als Kriterium wurde – mangels weiterer allgemeiner Möglichkeiten – folgendes gewählt:

> Falls ein Satz aus mehr als zwei Analyseeinheiten (d.h. Subsätzen oder Teilsubsätzen 82/6 bestand – in der Regel ein Hinweis darauf, daß eine komplexe Struktur, etwa eine Einbettung, vorliegt – und darüberhinaus die Funktion (Wortklasse) der künstlich mehrdeutigen unbekannten Wortform nur in Abhängigkeit von einer Lösungsmöglichkeit eines noch nicht gelösten (natürlichen oder künstlichen) Homographen ermittelt werden konnte 82/7, wird auf 'unsichere' Lösung erkannt. 82/8 <S. 83>

Das hier gewählte Kriterium hat sich wenig bewährt – das kann bereits im Vorgriff auf die Testergebnisse festgehalten werden. Der Grund dafür lag jedoch nicht im Bewertungsmaßstab selbst, die als 'unsicher' apostrophierten Lösungen waren in der Regel wirklich unsicher, also auch oft falsch: es gab darüberhinaus jedoch viele falsche Lösungen, die als ‚sicher' gewertet wurden, etwa weil sie in den

entsprechenden Homographenreduktionsprogrammen sehr rasch (zumindest im ersten Durchlauf) gelöst wurden, wodurch bei der Konzeption der Saarbrücker Analyse keine Handhabe mehr gegeben war, den Sicherheitsgrad der Lösung zu ermitteln. 84/1 Dies liegt mit darin begründet, daß für unbekannte Wörter – mangels eines speziellen Lösungsprogramms – etwa die gleichen Zuordnungsüberprüfungen (z.B. Kasus-Numerus-Kongruenz) durchgeführt wurden, wie sie bei natürlichen Homographen; die hier nicht die vielfältigen theoretisch möglichen, sondern die wenigen konkret vorhandenen Informationen besitzen, durchaus zu Wortklassenausschlüssen führen. 84/2

Die Merkmale 'unsicher' / 'sicher' wurden bei der Informationsverknüpfung im Lexikon durch eine einfache Regel folgendermaßen in Relation zueinander gesetzt:

> Tritt eine 'unsichere' Lösung für eine Wortform in bezug auf eine Wortklasse zweimal auf, so wird sie durch eine 'sichere' ersetzt (die unsicheren werden gelöscht).

Auf diese Weise werden etwa doppelt sö viele unsichere wie sichere Lösungen benötigt, um einen Schwellenwert zu erreichen. Eine unsichere Lösung ist also dahingehend betrachtet halb so viel wert wie eine sichere. Sie ist jedoch gar nichts wert, wenn ein Schwellenwert für eine andere Wortklasse erreicht wird und nur eine einzige unsichere Lösung für die betreffende Wortklasse vorliegt. <S. 84>

Daraus ergibt sich die Möglichkeit, ein Ergebnis einer Satzanalyse im Nachhinein zu korrigieren, sofern die Analyse (und Wörterbuchsuche) für diejenigen Sätze wiederholt wird, in denen eine später gelöschte und schließlich nach Erreichen der Kannphase 'verbotene' (gesperrte) Wortklassifizierung erfolgte. 84/1

4.2.2 Die quasimorphologische Präklassifikation

Die unbekannten Textwortformen werden – wie bereits angedeutet – keineswegs mit einer Mehrdeutigkeitsangabe ausgestattet, die alle nicht vollständig im Lexikon verzeich.. neten Wortklassen umfassen würde. Vielmehr wird über eine Analyse der Endgrapheme eine erste Einschränkung der Deutigkeiten vorgenommen. Ehe jedoch auf die dabei zu beachtenden Fragen eingegangen werden soll, sei zunächst die Liste der Endungen mit den daraus resultierenden Reduktionen angeführt. 84/2

LISTE DER KÜNSTLICHEN MEHRDEUTIGKEITEN

Typ 85/1	Endung	HO 85/2	K 85/3	D 85/4	Mehrdeutigkeit 85/5
1	IERT	8	3	-	ADV VRB PTZ
2	IEREN	4	4	-	VRB INF SUB
3	ER	46	-	4	ADJ ADV SUB
4	ES	5	-	5	ADJ SUB
5	EN	14.4	4	2	ADJ ADV VRB INF SUB
6	E	14.4	1	1	ADJ ADV VRB SUB
7	N	14.1	4	-	ADV VRB INF SUB
8	A,I,0,U	55	-	-	SUB 85/6
9	ST	10	2	-	ADV VRB PTZ SUB
10	T	10	3	-	ADV VRB PTZ SUB
11	EM	5	-	3	ADJ SUB
12	HAFTER85/7	20	-	4	ADJ ADV

13 (Rest) 85/8	11	-	-	ADV SUB
nach Verknüpfung -	7			ADJ VRB INF SUB
außerdem noch	4			SUB VRB INF
mögliche redu-	9			SUB PTZ ADV
zierte Homo graphen	6			VRB INF ADJ
	14.3			VRB INF ADJ ADV
	12			VRB INF ADV
	8			VRB PTZ ADV
	13			ADJ ADV
	2			VRB INF (SUB) 85/9
	45			PTZ ADV

<S. 86>

Es steht außer Frage, daß sich dieses hier angeführte Präklassifikationssystem noch verbessern ließe. Beispielsweise könnte der Typ 7 (-N) dahingehend eingeschränkt werden, daß als vorletztes Graphem ein R bzw. L auftreten muß, wenn die Lösung VRB/INF noch möglich sein soll; ähnliches gilt für die Typen 9 (-ST) und 10 (-T): die Lösung PTZ ist nur erlaubt, wenn am Wortanfang die Graphemfolgen GE- bzw. VER-, ZER-, UEBER-, ..., also eines der möglichen Präfixe, vorhanden sind; bei dem Typ 6 (-E) ließe sich die ADV-Möglichkeit ausklammern, falls alle in der Grundform auf -E endenden Adjektiv/Adverbien im Ausnahmelexikon verzeichnet worden wären (diese Bildungsmöglichkeit des Adverbs ist heute nicht mehr gegeben, d.h. die Liste dieser Wörter (WEISE, BEIGE ...) ist abgeschlossen). Eine derartige theoretische Annahme wurde übrigens bei dem Typ 5 (-EN) gemacht: Da die Starken Verben im PTZ II (unflektiert) auf -EN enden (GESUNKEN, VERSTANDEN, GEGESSEN ...), hätte hier eine weitere Deutigkeit PTZ angenommen werden müssen, d.h. diese sowieso schon am wenigsten reduzierbare Gruppe (es sind die Deutigkeiten ADJ ADV VRB INF SUB (NAM) zugelassen) wäre in bezug auf die noch möglichen Wortklassen überhaupt nicht mehr einschränkbar gewesen. Daher wurde angenommen, daß alle Partizipien der Starken Verben im Lexikon verzeichnet seien. 86/1

Es geht hier jedoch weniger darum, ein vollendetes Präklassifikationssystem anzugeben, als den Weg dazu zu zeigen. So könnten etwa morphologische Untersuchungen des Wortkörpers weitere Einschränkungen bringen (in unserem Falle würde etwa SCHOEN, falls es ein 'unbekanntes' Wort darstellte, über die vermeintliche Endung EN in die Gruppe 5 eingeordnet, eine kleine weitere morphologische Analyse hätte es aber der Gruppe 13 (ADV/SUB) zuordnen können); eine Wortlängenuntersuchung, etwa verknüpft mit Präfigierungsregeln, würde bei vielen überlangen Wörtern, die hier noch mit <S. 87> der Deutigkeit VRB versehen sein könnten, entsprechende Einschränkungen vornehmen; es wäre etwa auch zu prüfen, inwieweit das von GEENS fürs Englische verwendete Verfahren, das anhand der Graphemstruktur eine Wortklassenerkennung versucht, aufs Deutsche übertragbar wäre, d.h. inwieweit Graphemkombinationen im Wortinnern sichere Vereindeutigungen von Wortfunktionen oder zumindest sichere Wortklassenreduktionen zulassen. 87/1 Weitere Einschränkungsmöglichkeiten, die zweifelsohne zu einer Verbesserung der Reduktionsergebnisse beitragen können, ergeben sich für die Genus-, Numerus- und Kasusangaben. Eine Ausnahmeliste hätte es in manchen Fällen ermöglicht, die Deklinationsangaben der unbekannten Wörter weiter einzuschränken, als es für dieses Modell geschehen ist: Bei den Substantivs werden beispielsweise ohne Einschränkung alle Kasus, Numeri und Genera

zugelassen. Das hier vorgelegte Reduktionsverfahren trägt also auch in seinem Umfang und seinem erstrebten Feinheitsgrad dem Modellcharakter des gesamten Algorithmus Rechnung.

4.2.3 Die Auflösung der künstlichen Mehrdeutigkeiten

Das Problem der Erkennung, d.h. der Klassifikation 'unbekannter' Wörter, ist keineswegs neu. In den USA waren schon in den sechziger Jahren entsprechende Verfahren entwickelt worden – ebenfalls in Modellform –, 87/2 in denen vor allem automatische Wortklassenerkennungen versucht wurden, also ebensolche Reduktionsergebnisse erzielt werden sollten, wie sie das hier vorzustellende Modell aufzeigen will. Als Beispiel sei das WISSYN-Projekt herangezogen 87/3, das stellvertretend für alle mir dazu bekannten Verfahren kurz erläutert werden soll:

Es wird zunächst – wie kaum anders denkbar – ein Lexikon zugrundegelegt, das, soweit ich erkennen kann – in etwa <S. 88> die Elemente enthält, die sich aus unserer Forderung für die erste Gruppe des Basislexikons ergeben, obwohl es zugleich ein Lexikon der häufigsten Wortformen zu sein scheint; es enthält im wesentlichen Funktionswörter wie Konjunktionen und Präpositionen, auch Hilfsverben und häufige Verben wie THINK, REQUEST sind verzeichnet. Die nicht im Lexikon vorhandenen Textwörter werden mittels einer morphologischen Analyse anhand einer umfangreichen Endungsliste (70 'Endungen' wie ARY, HOOD, BILITY ...) zu klassifizieren versucht; auch hier eine deutliche Parallelität zu dem hier vorgeführten Modell. 88/1 Der wesentliche und m.E. entscheidende Unterschied dieser Verfahren zu dem Saarbrücker Analyseverfahren und damit auch zu dem zu schildernden Modell liegt in der Behandlung der nach der morphologischen Analyse noch verbleibenden unbekannten Wörter: In allen Projekten wird – fürs Englische sicherlich praktisch legitimer, da die Wortstellung fester und die Diskontinuität seltener ist – ein statistisches Wortstellungsmodell zugrundegelegt, das die Wortklasse eines unbekannten Wortes allein aufgrund seiner Stellung zu den es umgebenden (der Wortklasse nach bekannten) Wörtern zu ermitteln versucht. Der Erfolg ist wenn man den Verfassern Glauben schenken darf – beachtlich 88/2, das Verfahren in der Regel sehr ausgefeilt: verschiedene Stellungswahrscheinlichkeiten (...Muster) werden gegeneinander abgewogen, und es soll nicht verschwiegen werden, daß die Fehlerquote geringer ist als bei dem in Saarbrücken verwendeten, im folgenden zu skizzierenden Erkennungsverfahren. Das mag allerdings an der Struktur der Testsätze liegen, die hier m.E. wesentlich komplexer gewesen ist. <S. 89>

Die Schwäche dieser auf rein statistischen Methoden aufbauenden Verfahren, die durchaus praktikabel sein mögen, liegt m.E. in der Anschauung begründet, daß 'Ausnahmen' nicht linguistische Sonderfälle, sondern statistisch weniger ins Gewicht fallende Gegebenheiten sind, die folglich nicht erfaßt zu werden brauchen. Die Argumentation ist also rein ökonomischer Natur und nicht-linguistisch. Hier stellt sich die generelle Frage, inwieweit sich die (angewandte, auf die Performanz bezogene) Linguistik der Statistik bedienen darf.

Ohne Zweifel kommt der Statistik und der statistischen Wahrscheinlichkeit eine wichtige Rolle zu im Hinblick auf die Strategie des Vorgehens, also etwa im Zusammenhang mit der Geschwindigkeit, mit der ein Ergebnis erreicht werden kann. Ein weiteres wichtiges Argument für die Anwendung statistischer Methoden in der 'angewandten' Linguistik ist die Frage nach der Akzeptabilität einer Struktur, obgleich man hier sehr schnell zur Frage einer 'repräsentativen' Textauswahl

gelangt: hier kommt man in der Regel mittels intuitiver Kompetenzentscheidungen anhand konstruierter Beispiele ebenso weit, denn letztlich ist es die Kompetenz eines Sprecher/Schreibers, die Strukturen wie DER DEN ALTEN, EINEN DURCH DEN REGEN AUFGEWEICHTEN STRAUSS GELBER, FRISCH GEPFLÜCKTER ROSEN TRAGENDEN MANN SICHER GELEITENDE SCHUTZENGEL als akzeptabel anzusehen und sie somit zu äußern oder als inakzeptabel zu betrachten und sie also in akzeptable aufzulösen.

Wichtig ist die Statistik m.E. auch im Zusammanhang mit der Beschreibung eines Sprachsystems, dergestalt, daß den (seltenen) Ausnahmen nicht über Gebühr Raum gewidmet wird – vor allem, wenn es etwa wie im Fremdsprachehuntericht darum geht, das wesentliche von der Sprache, also ein Basiswissen, zu vermitteln. 89/1 <S. 90>

Für unseren Zusammenhang entscheidend ist jedoch die Frage, ob bei der Beschreibung der Kompetenz eines Sprecher/Hörers die Statistik dazu führt, daß die 'Ausnahmen' als nicht mehr zum Sprachsystem gehörig ausgeklammert werden, wie es notwendig bei den geschilderten Projekten geschehen muß. Ein an der Kompetenz orientiertes Grammatiksystem – und dies gilt m.E. auch für die maschinelle 'Grammatik', den Parser – darf in dieser Hinsicht keine statistischen Gesichtspunkte in Erwägung ziehen; d.h. die 'Ausnahmen' sind ebenso zu betrachten und zu beachten wie durch die Statistik im allgemeinen erfaßten Regeln, nämlich als Teile des Systems.

Dieser Gesichtspunkt ist dem Saarbrücker Analysesystem zugrundegelegt, ohne daß es allerdings konkret hätte verwirklicht werden können. 90/1 Die (natürliche Wortklassenmehrdeutigkeit wird dabei anhand kontextsensitiver Regeln aufgelöst. Dazu gehören natürlich Untersuchungen der den Homographen umgebenden Wortklassen (auf denen allein die bisher entwickelten statistischen Verfahren aufbauen), dazu kommen, falls möglich, Kongruenzüberprüfungen, d.h. Kontrollen auf Genus-, Numerus- und Kasusübereinstimmung bzw. Nichtübereinstimmung; wichtig ist auch die Stellung des Wortes oder der Wortgruppe im (Sub-)Satz (Satzanfang, Satzende ...) und das den Satztyp (Hauptsatz, Nebensatz) klassifizierende und damit auf die Wortstellung hinweisende funktionale Inventar wie Satzzeichen und Konjunktionen. Das I:ösungsergebnis wird im allgemeinen aufgrund einer Kombination der hier angedeuteten Regeln ermittelt.

Der Lösungsweg des Homographen hängt prinzipiell von den Wortklassen ab, die als Lösung in Frage kommen. Die Reduktionsprogramme sind also auf den Homographen zugeschnitten. <S. 91> Allerdings waren für die über 50 bisher bearbeiteten Homographentypen 91/1 nicht ebensoviele Lösungsalgorithmen nötig, da die Auflösung auch durch Koordination bereits für andere Typen erstellter Programme geschehen konnte, sofern deren Lösungen eine echte Teilmenge der Lösungen dieses Homographentyps darstellten.

Es wurde schon erwähnt, daß alle Mehrdeutigkeiten der 'künstlichen' Homographen bereits auch bei natürlichen Homographen auftreten. Es lag daher nahe, die entsprechenden Strategien und Regeln auch hierfür einzusetzen. Dabei wurden für diesen Fall keine besonderen Routinen vorgesehen; der Algorithmus machte also keinen Unterschied zwischen einem künstlichen oder natürlichen Homographen. Dies konnte als 'Nachteil' zuungunsten des 'künstlichen' Homographen nur dort wirksam werden, wo Entscheidungen allgemein aufgrund von Kongruenzäbfragen getroffen wurden, da die unbekannten Wörter – besonders die Substantive – hier ein

Zuviel an Information anboten. Evtl. wurde dieser Nachteil auch bei der weiteren Analyse wirksam, die dann allerdings die Wortklassenentscheidung nicht mehr beeinflußte, wenn etwa 'das Finite Verb (ein entsprechend gelöstes unbekanntes Wort) Dativ- und Akkusativ-Rektion hatte und entsprechende Objekte verlangte.

4.2.4 Der Analysezyklus

Es wurde schon darauf hingewiesen, daß das Modell zur Klassifikation unbekannter Wörter in das Saarbrücker Analysesystem integriert ist. Teile davon – etwa die morphologische und die quasimorphologische Präklassifikation sowie die Klassifikation durch den Analysealgorithmus selbst sind mittlerweile feste Bestandteile des Systems; das Kurzzeitlexikon kann wahlweise in den Analyseprozeß aufgenommen oder aber auch ausgeschlossen werden. Zum besseren Verständnis sei hier der gesamte Analysezyklus <S. 92> in seinem technischen Ablauf überblickartig dargestellt:

a) *Textaufnahme:* Befinden sich die zu analysierenden Daten nicht bereits auf Magnetband (wie es etwa bei den 11.000 Sätzen der rde- und FAZ-Corpora der Fall ist), so werden sie über 5-Kanal-Lochstreifen (Fernschreibcode) 92/1 eingelesen; der fortlaufende Text wird dabei automatisch in Sätze geteilt, 92/2 die eine Nummer erhalten, und auf Magnetband entsprechend aufgezeichnet.

b) *Satzvorbereitung mit Wörterbuchsuche:* Anschließend wird der Text wieder in den Computer geladen und im Kernspeicher alphabetisch anhand der Wortformen sortiert, wobei jedem Wort seine Satznummer und die Wortnummer innerhalb des Satzes zugefügt werden, um es später wieder in Textreihenfolge ordnen zu können. Jedem Wort werden zugleich alle ihm unmittelbar folgenden Satzzeichen mitgegeben.

   Für diese alphabetisch sortierten Wortformen wird jetzt die Wörterbuchsuche durchgeführt. Befindet sich das Wort im 'normalen' Wortformenlexikon, so werden ihm die dort verzeichneten grammatischen Informationen zugeordnet. Ist die nicht der Fall, wird – falls eine entsprechende Wahl getroffen ist – das Kurzzeitlexikon nachgeschlagen und geprüft, ob das Wort dort bereits aufgenommen ist. Ist dies der Fall, so erfolgt in keinem Falle eine weitere morphologische' Analyse, obgleich sich noch zwei Möglichkeiten ergeben: Entweder ist (sind) die Wortklasse(n) bereits eindeutig bestimmt (= mehr als 7 sichere Lösungen); dann wird das entsprechende Ergebnis der Textwortform zugeordnet; andernfalls wird der Homographentyp anhand der noch möglichen Lösungen ermittelt und beim Textwort entsprechend verzeichnet.

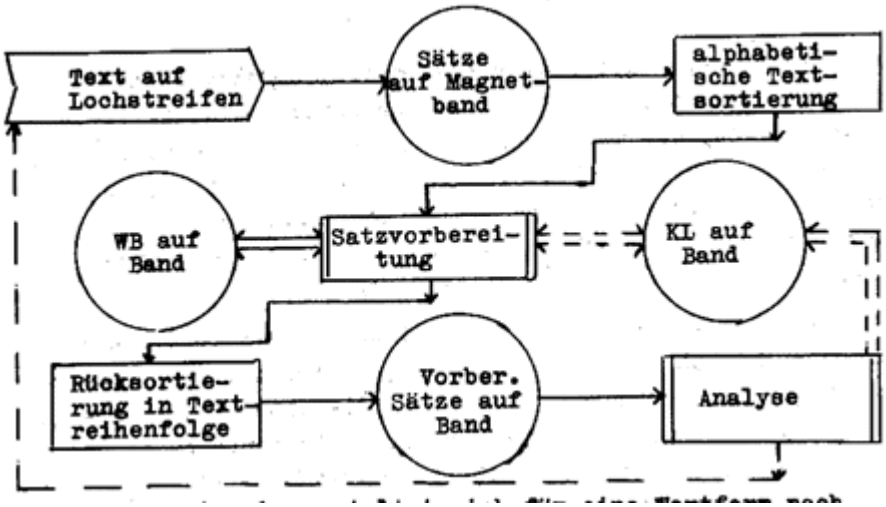
   Befand sich das Wort nicht im Kurzzeitlexikon oder sollte das Kurzzeitlexikon nicht erstellt bzw. abgefragt werden, so wird die morphologische bzw. quasimorphologische <S. 93> Analyse durchgeführt (beide Listen sind physikalisch gesehen vereint). Die sich dabei ergebenden Klassifikationen werden der Textwortform zugeordnet. Wurde eine Wortklassenmehrdeutigkeit festgestellt, so wird, falls gewünscht, wieder das Kurzzeitlexikon angesprochen, das Wort dort im Wortlaut aufgenommen und die entsprechenden Möglichkeiten und werte festgehalten.

   Sind allen Textwörtern auf eine dieser Weisen ihre grammatischen ;angaben zugeordnet worden, so wird der Text anhand der Satz- und Wortnummern

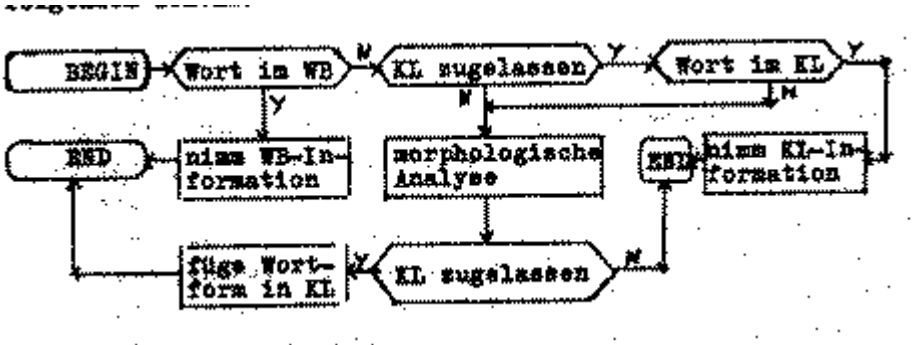
wieder in die alte Textreihenfolge zurücksortiert, wobei die durch die Wörterbuohsuche gewonnenen Informationen natürlich mitgeführt werden.

Die so vorbereiteten Sätze werden auf Magnetband•zwischengespeichert.

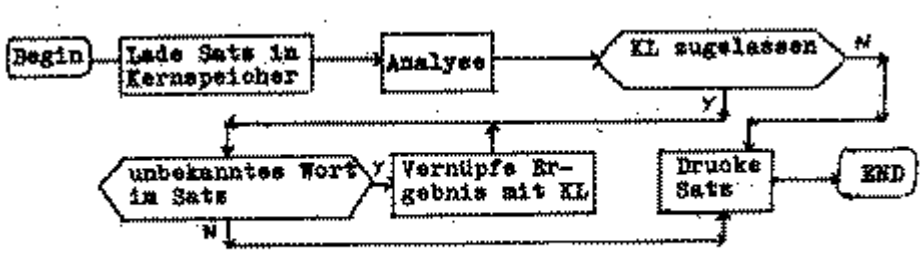
c) *Analyse:* Anschließend werden die Texte Satz für Satz bearbeitet. Jeweils ein Satz wird also in den Speicher geladen und grammatisch analysiert. Das Ergebnis wird auf einem Schnelldrucker ausgegeben. Falls das Kurzzeitlexikon eingeschaltet ist, werden alle unbekannten Wörter (eine entsprechende Information ist bei der Wörterbuchsuche festgehalten worden) in diesem Lexikon aufgesucht; das Ergebnis der Satzanalyse wird mit den dort verzeichneten Informationen entsprechend den oben aufgezeigten Regeln verknüpft. Sind alle Sätze auf diese Weise analysiert und bearbeitet, ist ein Analysezyklus geschlossen, d.h. der Rechner ist wieder bereit, neue Sätze (über Lochstreifen) einzulesen und die Prozedur zu wiederholen.
Zur Veranschaulichung eine Skizze des Analysezyklus 93/1: <S. 94>



Die Wörterbuchsuche gestaltet sich für eine Wortform nach folgendem Schema:



Das Verknüpfen der Analyseergebnisse mit dem Kurzzeitlexikon läßt. sich folgendermaßen veranschaulichen:



<S. 95>

4.3 Ein Test des Modells

Sicherlich gibt es eine Reihe von Fällen in der Linguistik, bei denen etwa durch Definitionen ein sprachlicher Bereich fest und exakt abgrenzbar ist und ein Regelsystem entwickelt werden kann, das in sich schlüssig ist und sogar in mathematischer Formelsprache dargestellt oder bewiesen werden kann, also den Testfall gar nicht benötigt. Das hier vorgelegte Modell hat nicht diese klare Einsichtigkeit vorzuweisen, es kann sich kaum einem theoretischen Beweis stellen. Wenn auch auf theoretischen Vorstellungen aufbauend, ist dieses Modell doch performanz-bezogen, also anwendungsorientiert. Es lag daher der Versuch nahe, die Verwendbarkeit der Methode anhand empirischer Untersuchungen nachzuweisen. Die Entscheidung für einen Testversuch fiel um so leichter, als das konkrete Modell ja 'gebrauchsfertig' vorlag.

Es soll dabei nicht darum gehen, einen wirklich schlüssigen Beweis dafür zu erbringen, daß damit ein oder gar der einzig sichere Ausweg aus dem Dilemma der Diskrepanz zwischen einem deskriptiven Grammatiksystem, das notwendig unvollständig sein muß, und einer Idealgrammatik gefunden worden sei. Dafür ist dieses Modell – auch im Bereich des Lexikons – vielleicht zu grob, die Analysägrammatik zu inadäquat und das Testmaterial womöglich nicht ausreichend. Es wird schon genügen, wenn sich damit eine Möglichkeit ergibt, die Kluft zwischen einem nie erreichbaren vollständigen Lexikon für die maschinelle Sprachanalyse und einem realiter konstruierbaren Maschinenwörterbuch auf dem hier vorgeschlagenen Weg zu überwinden.

4.3.1 Ziel und Vorgehensweise

Der Test sollte zunächst dazu dienen, die Brauchbarkeit einer derartigen Methode zu beweisen. Daher mußte eine ausreichende Materialmenge zugrundegelegt werden, die es erlauben konnte, die Ergebnisse mit einiger Aussagekraft statistisch auszuwerten. Da für diese Auswertung andererseits <S. 96> eine große Zahl von Einzelergebnissen von Bedeutung war und auch möglichst viele Zwischenergebnisse – vor allem im Hinblick auf die einzelnen Stadien des Kurzzeitlexikons festgehalten werden sollten, schien es sinnvoll, die Textmenge dennoch überschaubar zu halten.
Damit schied die Möglichkeit aus, einen 'Random'-Zugriff zu irgendeinem beliebigen computerzugänglichen Text durchzuführen, es also dabei völlig dem Zufall zu überlassen, welche Wörter und Sätze analysiert werden sollten. Die für den Test notwendigen Ergebnisse hätten eine Analyse sehr großer Textmengen vorausgesetzt, und es wäre dennoch zu fragen gewesen, inwieweit dann diese Textmenge als repräsentativ betrachtet werden könnte.

Für die Art der 'Auswahl' des Testmaterials gab schließlich der zweite Gesichtspunkt den Ausschlag, unter dem der Test durchgeführt wurde. Es galt, mithilfe des Modells praktische Erfahrungen zu sammeln und Möglichkeiten zur Verbesserung oder Veränderung des Verfahrens zu finden, d.h. ebtl. vorhandene Schwächen aufzuzeigen. Daher sollte die Auswahl der Wörter nicht völlig dem Zufall überlassen bleiben; andererseits sollte die Testbedingungen – also ein vom Verfahren unbeeinflußtes Material – gewahrt bleiben, um den Anspruch einer Analyse beliebiger Sätze aufrechterhalten zu können. Zugleich sollten Erfahrungen mit der Verwendung des Kurzzeitlexikons und den dabei vorgenommenen Verknüpfungen gesammelt werden. Daher wurde folgendermaßen verfahren:

Einige 'native Speakers' wurden ersucht, mithilfe vorgegebener Wortformen beliebige Sätze zu bilden. Diese Wortformen waren nicht im maschinellen Lexikon der Syntaxanalyse vorhanden, also 'unbekannt'. Um zu einer endgültigen Klassifikation einer derartig unbekannten Wortform im Kurzzeitlexikon zu gelangen, waren acht 'sichere' Lösungen nötig. Daher mußte im gesamten Testmaterial jede dieser Wortformen mindestens achtmal vorkommen. Um hierbei eine <S. 97> möglichst gestreute Kontextsituation zu gewährleisten, sollte jede 'Testperson' jeweils nur einen Satz mit einem der unbekannten Wörter konstruieren, die Wahl der übrigen Wörter des Satzes war völlig freigestellt. 97/1 Als Testpersonen wurden vorwiegend (acht von insgesamt elf 'native Speakers') Mitarbeiter des Forschungsprojekts 'Automatische Lemmatisierung' herangezogen, die teilweise auch mit dem Verfahren der Syntaxanalyse vertraut sind. Hier sollte die Gefahr, daß damit die 'Unbefangenheit' bei der Satzkonstruktion verlorengehen konnte, dadurch aufgewogen werden, daß etwas 'härtere' Testbedingungen zustande kamen, da ein unvoreingenommener, im bewußten Konstruieren von Sätzen ungeübter native Speaker eher dazu zu neigen scheint, 'einfachere' Sätze zu bilden. 97/2 Nähere Einzelheiten sind der im folgenden wortgetreu wiedergegebenen Anleitung zur Satzkonstruktion zu entnehmen:

> ‚Bitte bilden Sie unter Verwendung der folgenden – nicht im Syntax-Wörterbuch vorhandenen – Wortformen jeweils mindestens einen Satz bis zu 32 Wörtern Länge. Die mit * versehenen Wörter sind in ihrer syntaktischen Funktion mehrdeutig. Die Wahl der übrigen Wörter des Satzes ist freigestellt, doch sollen es keine ungewöhnlichen oder ungebräuchlichen Wortformen sein. In der Mischung der Satzstrukturen soll in etwa das durchschnittliche Strukturverhältnis im Deutschen zum Ausdruck kommen (also nicht nur 32-Wort-Sätze, nicht in jedem Satz eine Diskontinuität, auch nicht nur Einfachsätze oder einfache Verbal- bzw. Nominalgruppen). <S. 98>
>
> Die Wortformen:
> *GELBER BUNDESPOLITIKER ANALOGER; DUNKELBRAUNES MACHTKOMPLEXES; DUNKELGRUENEN *BEANSTANDETEN KOSAKEN *BEDECKEN; DUNKELBRAUN *ANLÄCHELN; GRINST EINGELOEST PEST FELSENFEST; AXT *BEGAFFT KRAXELT; DUNKELBRAUNEM PHONEM; ANLAUFS BERGAB ABDRUCK; DUNKELBRAUNE ANGSTHASE VERBEUGE. Es sind also mindestens 26 Sätze zu bilden.'

Nähere Einzelheiten wurden den Testpersonen nicht mitgeteilt. Die Testprogramme waren ihnen nicht bekannt.98/1

Da zu vermuten war, daß die konstruierten Sätze in einigen Punkten (Anteil der unbekannten Wörter am Gesamttext, Durchschnittslänge der Sätze, Komplexität der Satzstrukturen) kein völlig zuverlässiges Bild von der Qualität der Klassifizierungsergebnisse vermitteln könnten, wurde darüberhinaus ein kleines Corpus nach Zufälligkeitsgesichtspunkten ausgesuchter Zeitungs- und Zeitschriftenartikel zusammengestellt und zu Kontrollzwecken analysiert. Diese Analyseergebnisse sind in Kap. 4.3.3.3 beschrieben.

4.3.2 Das Testmaterial

Die Testsätze mit vorgegebenen Wortformen wurden in der Reihenfolge der abgelieferten Beiträge zu einem Corpus zusammengestellt und auf Lochstreifen übertragen. Es waren insgesamt 282 Sätze konstruiert worden, 98/2 die sich den Testpersonen 98/3 folgendermaßen zuordnen lassen:

Testperson	*Sätze*	*Satzanzahl*	*Wortzahl*	*Durchschnitt*
1	1 – 27	27	431	18,6
2	28 – 53	26	174	6,7 <S. 99>
3	54 – 79	26	169	6,5
4	80 – 105	26	360	13,6
5	106 – 131	26	299	11,5
6	132 – 158	26 99/1	266	10,2
7	159 – 184	26	374	14,4
8	185 – 210	26	352	13,6
9	211 – 235	25	294	11,8
10	236 – 261	26	279	10,7
11	262 – 283	22	180	89,2
Gesamt		282	3178	11,2

In den 282 Sätzen sind also insgesamt 3178 laufende Wortformen enthalten, daraus ergibt sich ein Durchschnittswert von 11,2 Wörtern je Satz. Von diesem Mittelwert wichen die Testpersonen 2 und 3 (nach unten) sowie 1, 7 und 8 (nach oben) augenfällig ab. 21 dieser Sätze, also etwa 7,4 %, haben weniger als 5 Wörter; 14 Sätze, also ca. 5 % haben mehr als 24 Wörter; der häufigste Satz ist der 8-Wort-Satz mit 36 Belegen. 99/2 Vergleicht man diese Werte mit den in Saarbrücken ermittelten Daten zu populärwissenschaftlicher Prosa und zu Zeitungstexten, so scheinen die Testsätze -. gemessen an der Satzlänge – weniger repräsentativ zu sein. 99/3 Unter diesem – wenn auch m.E. nicht sehr schwerwiegenden – Vorbehalt sind daher auch die Vergleiche zu sehen, die im Anschluß an die Beschreibung der Testresultate zu den Ergebnissen der Syntaxanalyse gezogen werden. Einen eingehenderen Vergleich zu dem Corpus der Syntaxanalyse durchzuführen – etwa im Hinblick auf die Komplexität der Sätze, die sich in Verschachtelungen und Einbettungen, in Anreihungen, in <S. 100> Nominal- und Verbalstrukturen äußert, würde, sofern dazu bereits Material vorliegt – ins Uferlose führen.

Dennoch scheint es angebracht, wenigstens einen Eindruck von der Komplexität der Testsätze zu vermitteln, ohne mittels der wenigen Beispiele ein festes Urteil erreichen zu wollen. Ein Blick auf einige (auch kurze) Sätze zeigt bereits, daß eine komplexe Struktur – vor allem bei diesen oft gerade daraufhin angelegten, ausgeklügelten Testsätzen – durchaus üblich ist: So stehen einigermaßen ungewöhnlich strukturierte kurze Sätze wie der 'Hauptsatz'

40 EINGELÖST WURDE SEIN VERSPRECHEN NIE,1

auch 'ungrammatische' Sätze wie

43 IM WALD DIE AXT ERSETZT DIE FLINTE.

oder das 4-Wort-Satzgefüge

39 WER GRINST, TOETET NICHT.

neben durchaus gewöhnlichen kurzen Sätzen wie

44 BEGAFFT IHN NUR!

oder

48 BERGAB GEHT ES SCHNELLER.;

einfach strukturierte lange Sätze wie

15 NACHDEM ER AUSFUERLICH LIEBER DIE VERHEERENDEN FOLGEN DER PEST GESPROCHEN HATTE, STIMMTEN ALLE ANWESENDEN DEM ANTRAG ZU, DER POLIZEI MEHR RECHTE GEGENÜBER DEMONSTRIERENDEN STUDENTEN ZU GEBEN!

stehen neben sophistisch anmutenden, die Grenze der Akzeptabilität erreichenden Sätzen wie

170 WEIL, WER, WENN ER EINEN, DER IHM UEBERGEORDNET IST, IN UNTERHOSEN SIEHT, GRINST ODER BLOED LACHT, SPINNT, IST IHM ZU VERZEIHEN.

Bei einigen Sätzen ist also zu vermuten, daß sie nicht so <S. 101> `keimfrei', so unbefangen konstruiert sind, sondern vielleicht unter der Fragestellung, was wohl der Computer damit anfangen mag.

So etwa auch

275 PEST UND BUDA LIEGEN EINANDER GEGENÜBER. 101/1

oder der metasprachliche Satz

42 BILDEN SIE NOCH 1INEN SATZ MIT ' FELSENFEST '.

Alle diese Sätze wurden ohne weitere Präedition (also auch unter Verzicht auf die Kennzeichnung der Substantive durch Großschreibung) dem Test unterzogen.

4.3.3 Testergebnisse

In Abhängigkeit von den wichtigsten Komponenten des Tests lassen sich die Ergebnisse in drei Abschnitte gliedern:

a) die Auswertung der morphologischen Analyse;
b) die Ergebnisse der Syntaxanalyse (Homographieauflösung);
c) die Betrachtung des Kurzzeitlexikons.

Um auch vom Kurzzeitlexikon unbeeinflußte Analyseergebnisse zu erhalten, die geeignet waren, mit den Ergebnissen der Syntaxanalyse verglichen zu werden, wurden die elf Zyklen jeweils einmal ohne Verwendung und einmal unter Einbeziehung des Kurzzeitlexikons durchgeführt. In Abschnitt 4.3.3.2 werden die Ergebnisse der 'Analyse ohne Kurzzeitlexikon' behandelt werden, während die Resultate der 'Analyse unter Verwendung des Kurzzeitlexikons' im Abschnitt 4.3.3.3 besprochen werden.

4.3.3.1 Ergebnisse der morphologischen Analyse 101/2

Für den Test waren den Testpersonen keine Wörter vorgegeben worden, die hätten automatisch morphologisch eindeutig erkannt, d.h..deren Wortklasse hätte bestimmt werden können. Dennoch wurden 62 der 3178 laufenden Wortformen morphologisch klassifiziert. Dies bedeutet zwar nur einen Textanteil von <102> etwa 2 % am Gesamttext, doch wird dieses Verhältnis aufgewertet, wenn man es in Relation setzt zu den insgesamt unbekannten (laufenden) Wortformen. 561 oder 17,7 % aller Wörter des Textes waren 'unbekannt'; davon blieben also noch 15,8 % für die Analyse. Aber auch dieses Verhältnis ist – geprägt durch die notwendi£ unbekannten Wörter – noch keine gute Vergleichsbasis. Eine verläßlichere Zahl ergibt sich aus dem Verhältnis der nicht vorgegebenen Wörter im Hinblick auf die morphologische Analyse: 206 der 268 nicht vorgegebenen unbekannten Wörter, also 6,1 % des Gesamttextes, mußten über die syntaktische Analyse klassifiziert werden, für 62, also etwa ein Viertel der 268 Wörter, fügte die morphologische Analyse. Dabei waren 25 Wortformen als Adjektive, 6 als Adverbien, 26 als Substantive und 5 als Eigennamen bestimmt worden.

Die folgenden 6 dabei entstandenen Fehler (ca. 10 % der morphologisch klassifizierten Wortformen) sind näher zu untersuchen:

1. ELEGANZ (239) 102/1, über das Endgraphem Z als NAM (Name) klassifiziert: Eine vollständigere Suffixliste (-ANZ, ähnlich wie bei -ENZ ...) hätte hier den Fehler vermeiden helfen; vielleicht sollte das Endgraphem Z auch zu dem Homographenprogramm 55 (SUB/NAM) führen, also die Deutigkeit Substantiv noch offenlassen.

2. FRAUENBEKANNTSCHAFTEN (163): die gleiche Ursache wie bei ELEGANZ; hier wurde, da das Suffix -SCHAFT noch nicht in die Liste aufgenommen ist, über das Suffix -HAFT eine falsche Wortklassenbestimmung durchgeführt.

3. GESCHAEFTSGEBAREN (197), OHRFEIGE (14), POLIZEISTELLE (10l): Diese Fehler der morphologischen Analyse sind entweder nur über ein Kompositionserkennungsprogramm zu reduzieren, das für die Syntaxanalyse und dieses Verfahren hier nicht vorliegt oder es müßten alle Komposita zu GEBAREN, FEIGE oder STELLE als Ausnahmen verzeichnet sein, um eine <S. 103> derartige falsche Analyse zu verhindern.

4. ZISCHEN (218): auch hier eine ähnliche Möglichkeit, die falsche Zuordnung zu dem Suffix ISCH durch Aufnahme des Wortes in das Lexikon zu verhindern.

Eine weitere graphematische Untersuchung des Wortkörpers hätte ebenfalls dazu führen können, den Fehler zu vermeiden: ein Konsonant allein kann nie zusammen mit einem Suffix eine Ableitung bilden.

Die automatische Anwendung der 'Substantivierungsregel für Adjektive' des Analyseprogramms – für eine Fehlerkorrektur eigentlich nicht vorgesehen – brachte bei den morphologisch falsch analysierten Wörtern erstaunliche Ergebnisse: Die fünf fälschlich als ADJ klassifizierten Substantive FRAUENBEKANNTSCHAFTEN, GESCHAEFTSGEBAREN, OHRFEIGE, POLIZEISTELLE und ZISCHEN wurden allesamt in SUB umgewandelt; das gleiche geschah (richtigerweise) mit dem als Adjektiv klassifizierten Wort UNDEFINIERBARES und fälschlich mit dem Wort ELFTE 103/l. Das einzige Wort, das also (von ELFTE, das ja ADJ sein kann, abgesehen) nach der Satzanalyse wirklich falsch klassifiziert war, ist das Substantiv ELEGANZ. Wenn auch die oben erwähnten 'Korrekturen' nicht als zufriedenstellende Lösungen anzusehen sind, so hat sich die Brauchbarkeit einer morphologischen Analyse – zumindest für den syntaktischen Bereich – wohl qualitativ und quantitativ erwiesen.

4.3.3.2 Ergebnisse der Auflösung künstlicher Mehrdeutigkeiten

499 unbekannte Wortformen (= 223 verschiedene Wortformen, für die aber 499 automatische Klassifikationen durchgeführt wurden) sind mithilfe der syntaktischen Analyse klassifiziert worden; fast genau 80 % davon (398) sind richtig gelöst, d.h. im Satz der Wortklasse nach richtig klassifiziert <S. 104> worden. Aber auch hier genügen diese an den Anfang gestellten Orientierungsdaten natürlich nicht, eine hinreichende Aussage über die Brauchbarkeit des Verfahrens daraus abzuleiten. Entscheidend ist die Interpretation dieser allgemeinen Aussage mittels der im einzelnen gewonnenen Ergebnisse.

Ich will versuchen, das Material dahingehend zu beschreiben, daß in manchen Fällen auf die Tabellen des Anhangs verwiesen wird, während hier nur die groben Zusammenhänge oder auch in Auswahl Besonderheiten aufgezeigt werden sollen. Natürlich werden die statistischen Werte interpretiert werden müssen. Dabei kann man nicht immer der Gefahr einer subjektiven Beurteilung entgehen, zumal einige Rückschlüsse auf Ursachen oder Änderungsmöglichkeiten spekulativer Natur sein müssen.

Eine erste wichtige Beurteilungsgrundlage,.allerdings weniger für das hier vorgestellte Modell als für das Leistungsvermögen der Syntaxanalyse, ergibt sich aus einer Gegenüberstellung der Lösungsergebnisse der unbekannten Wörter, geordnet nach Wortklassen.

Wortklasse	*Belege*	*davon falsch gelöst (abs.)*	*falsch (rel.)*
SUB	245	45	18,4 %
ADJ	90	11	12,2 %
VRB	83	12	14,4 %
ADV	44	22	50,0 %
PTZ	25	3	12,0 %
INF	12	8	66,7%

Extreme Schwierigkeiten, die Wortklasse richtig zu ermitteln, bestehen bei INF- und ADV-Lösungen. Bei beiden Wortklassen scheint die Anzahl der Deutigkeiten eines

Homographen für die hohe Zahl falscher Lösungen mit verantwortlich zu sein. Während das Adverb in 14 von 18 Fällen, in denen die <S. 105> zweideutige Mehrdeutigkeit HO 11 (SUB/ADV) vorlag, richtig bestimmt werden konnte, wurde es in den 12 vierdeutigen Fällen von HO 10 (ADV/VRB/PTZ/SUB) nur dreimal richtig ermittelt. Ähnliches gilt für den Infinitiv, der bei dem vierdeutigen HO 14-n (ADV/SUB/VRB/INF) bzw. dem fünfdeutigen HO 14-en (ADJ/ADV/VRB/INF/SUB) bei insgesamt 11 möglichen Lösungen zu 8 falschen Ergebnissen führte. Es scheint sich hier also die aufgrund der Probe der Syntaxanalyse von 1969 105/1 aufgestellte These zu bestätigen, daß der Anteil richtiger Lösungen abhängt von der Anzahl der Lösungsmöglichkeiten eines Homographen.

Daß dies nicht in aller Absolutheit gilt, zeigt wiederum die Auflösung der Mehrdeutigkeiten PTZ/X: Genau genommen sind bis auf einen Fall (AXT in Satz 200) alle Partizipien richtig klassifiziert worden; in vier Fällen wurde eine Wortform fälschlich als PTZ gelöst (bei theoretisch gegebenen weiteren 82 möglichen falschen PTZ-Lösungen: dreimal hätte auf SUB, einmal auf VRB erkannt werden müssen; in zwei Fällen wurde eine Wortform nicht als PTZ gelöst, da diese Möglichkeit über die quasimorphologische Analyse ausgeschlossen worden war. 105/2

Für dieses Problem, daß eine Wortklasse bei gleichem HO-Typ in fast allen Fällen richtig aufgefunden wurde (nämlich PTZ), während eine andere, nämlich ADV, extrem schlecht erkannt wurde, gibt es zwei mögliche Erklärungen. Zunächst eine 'linguistische': Das Adverb ist im Deutschen verhältnismäßig stellungsfrei, es kann an nahezu allen Stellen stehen, an denen andere Wortklassen auftreten können. Dies bei der Erkennung unbekannter Wortformen zu berücksichtigen fällt besonders schwer, da ja im allgemeinen nicht nur ein mehrdeutiges Wort im Satz auftritt, während das Partizip II aufgrund einfacherer Stellungs- und sInstiger kontextsensitiver <106> Regeln (etwa muß ein Hilfsverb oder Modalverb auftreten) bestimmt werden kann. Die zweite Erklätungsmöglichkeit, die vor allem für die unverhältnismäßig hohe Fehlerquote bei der richtigen Erkennung des Infinitivs verantwortlich zu sein scheint, ist am technischen und linguistischen Zustand der Analyseprogramme (Homographenlösungsroutinen) orientiert. Nicht für alle Wortklassen bestand in der Kürze der zur Verfügung stehenden Zeit die Möglichkeit, sie mit der gleichen ausgiebigen Intensität zu erstellen oder zu testen, wie es etwa für die Mehrdeutigkeit DEMonstrativwort/RELativwort geschehen ist. 106/1

Die Ursachen für das schwache Abschneiden der beiden Wortklassen Adverb und Infinitiv in HO 14 bzw. HO 10 scheint also aus einer Mischung dieser aufgezeigten Möglichkeiten (komplexe Homographentypen, Stellungsfreiheit, mangelnde technische und linguistische Qualität des Parsers) erklärbar zu sein. 106/2

Betrachtet man die Wortklassen SUB, ADJ und VRB im allgemeinen Bezug auf ihre Lösungen oder Lösungsmöglichkeiten, so lassen sich kaum besondere Erkenntnisse damit verbinden. Interessant ist auch hier, daß von 45 falsch gelösten SUB 18 als ADV bestimmt wurden, während kein Adjektiv und nur zwei Verben fälschlich als ADV gelöst worden sind. Am wenigsten aufschlüsselbar sind die Lösungen bzw. Falschlösungen VRB/SUB, ADJ/SUB und ADJ/VRB, die gleichmäßig in beiden Richtungen streuen.

Interessant, wenn auch aufgrund der geringeren Belagzahlen statistisch nicht immer sehr ergiebig ist eine Untersuchung der Ergebnisse im Hinblick auf die einzelnen Homographentypen, da hierzu entsprechende Vergleichszahlen aus der Probeanalyse der 1475 rde-Sätze vorliegen. 106/3

Homogr.-typ	D107/1	*Test unbek. Wörter* Belege	richtig	rel.	*Probe Syntaxanalyse* Belege	richtig	rel.
5	II	58	52	90%	161	149	93%
10	V	107	86	80%	53	41	77%
11	III	64	58	91%	247	226	91%
14	V	213	157	73%	23	17	74%
46	III	47	36	77%	24	21	87%
*Ges.*107/2:		489	389	80%	518	454	88%

Es würde der Sache nicht gerecht werden, wenn man von den für diese fünf wichtigsten Homographentypen ermittelten Gesamtwerten richtiger Lösungen ausginge. Die Differenz von absolut 8 % (88 % der Syntaxanalyse gegenüber 80 % der Testsätze) wird wesentlich getragen von dem hohen absoluten Anteil der HO 11-Belege der Syntaxanalyse (247 gegenüber den 64 des Tests), die in beiden Fällen – relativ betrachtet – gleichermaßen gut analysiert und klassifiziert worden sind (jeweils 91 %). Nimmt man den Mittelwert der .gesamten Lösungen zueinander, d.h. geht man von gleichen absoluten Zahlenwerten für beide Gruppen aus, (als absoluter Wert wurde jeweils die Summe der beiden absoluten Ergehnisse für einen Homographen-Typ gewählt),so ergibt sich mit einer etwa 1,3 % besseren Analysequote der Syntaxanalyse relativ zu der Testanalyse kein allzu bedeutender Unterschied mehrb, die Ergebnisse der Probesätze von 1969, gewonnen an der Auflösung natürlicher Mehrdeutigkeiten, sind also nur um ein geringes besser als die Testergebnisse, gewonnen an der Auflösung künstlicher Mehrdeutigkeiten.3 Die Verschlechterung der Testergebnisse hat also vorwiegend quantitative Ursachen: die bei der Analyse natürlicher Mehrdeutigkeiten seltene und wenig <S. 108> umfangreiche Gruppe HO 14 tritt im Testversuch – also bei 'künstlichen' Mehrdeutigkeiten unbekannter Wörter ungleich häufiger auf, und gerade diese Mehrdeutigkeit kann nur um absolut 1 % schlechter gelöst werden als die entsprechende Deutigkeit natürlicher Wortformen (wohl zugleich eine Frage der Verknüpfungs- und Kongruenzregeln). Ähnlich verhält es sich bei den Homographentypen HO 44, HO 5 und HO 11, die quantitativ bedeutend zahlreicher sind als in dem Testmaterial; nur HO 10 macht hier eine Ausnahme, das um absolut gesehen 3 % besser gelöst wurde als in der Syntaxanalyse-Probe und auch quantitativ gesehen überwiegt.

Faßt man die bisher wichtigsten Ergebnisse in einen Satz, so kann man sagen, daß die Testergebnisse im Vergleich zur Untersuchung der Syntaxanalyse quantitativ um etwa 10 %, qualitativ dagegen um wenig mehr als 1 % abweichen. Für den Bereich der automatischen Klassifizierung unbekannter Wortformen nach der Wortklasse scheint die Aussagekraft dieses Tests und besonders des Vergleichs zur Probe der Syntaxanalyse auszureichen, diesen Weg als erfolgversprechend anzusehen. Daß eine Wortklassenerkennung erst den Anfang dieser Strecke darstellt, braucht wohl nicht weiter betont zu werden.

Ich will es mir hier versagen, ausführlicher auf weitere Einzelheiten und Sonderfälle der Analyseergebnisse einzugehen. Vielleicht noch einige Beispiele für eine durch

die Strategie des Saarbrücker Verfahrens hervorgerufene Fehlerursache, die sog. 'Fehlerfortpflanzung' 108/1; d.h. das Phänomen, daß eine fehlerhafte Homographenauflösung weitere fehlerhafte Lösungen anderer Homographen im Satz hervorruft. Solch ein Fall liegt etwa vor in

219 VOM STURM GEPEITSCHTE WOLKENFETZEN BEDECKEN DEN HIMMEL, ... ,

wo die unbekannten Wortformen GEPEITSCHTE, WOLKENFETZEN <S. 109> BEDECKEN (alle HO 14-) zu den dreifach falschen, sich gegenseitig bedingenden und formal-syntaktisch durchaus 'richtigen' 'Tortklassenfolgen VRB – ADJ – SUB führten, wodurch die formal wiederum korrekte Subsatz-Struktur
entstand: 109/2

VOM STURM	- Präpositionalgruppe
GEPEITSCHTE	- Prädikat
WOLKENFETZEN BEDECKEN	- Subjekt
DEN HIMMEL	- Akk-Objekt.

Eine derartige Struktur würde etwa der Satz VOM GROSSVATER ERWARTETEN BRAVE KINDER EIN GESCHENK besitzen. Eine Fehlerfortpflanzung stellt auch die Struktur ADV KON ADV (KOSAKEN UND TATAREN) des Satzes 192 dar, wo KOSAKEN (als unbekanntes Wort) fälschlich ADV wurde und in 'Anreihung' das ebenfalls nicht im Lexikon vorhanden Wort TATAREN gleichfalls zu ADV gelöst wurde. 109/3

Eine weitere 'Fehlerquelle' war bedingt durch das Fehlen einer Möglichkeit, echte formal-syntaktische Mehrdeutigkeiten als solche zu erkennen. Die Analyse bietet " wie schon erwähnt – in selchen Situationen stets nur eine wahrscheinlichere oder für wahrscheinlicher gehaltene Lösung an. Ein deutliches Beispiel dafür ist der 3-Wort-Satz

107 BRANDT IST BUNDESPOLITIKER.

Das unbekannte Wort BUNDESPOLITIKER war über die Graphem-Endung ER als 'mehrdeutig' ADJ/ADV/SUB präklassifiziert, die Analyse bot eine der beiden möglichen Syntaktischen Lösungen, nämlich ADV, an. Dies wäre etwa in dem Satz BRANDT IST KLUEGER 'richtig' gewesen. Ähnliches gilt für den Subsatz

87 VERSCHMITZT GRINST FRITZ AUS DEM FENSTER HERAUS UND ... .

Hier ist VERSCHMITZT (HO 10) fälschlich als SUB und nicht <S. 110> als ADV gelöst; daß syntaktisch gesehen ein SUB an dieser Stelle möglich gewesen wäre, zeigt der Satz ANGST GRINST FRITZ AUS DEN AUGEN ... . Eine formal richtige Lösung ist auch Satz 52 ANGSTHASE (ADV) NENNT MAN MISTEN (ADJ) EULENBAUM (SUB), nach dem Satzmuster WEISE NENNT MAN ERFAHRENE MAENNER. 110/1

Ähnliches gilt für die 'falschen' Substantivierungen. Ein Beispiel dafür ist etwa Satz

157 ICH WILL KEIN DUNKELGRUENES KLEID, LIEBER WAERE MIR EIN DUNKELBRAUNES.

wo die Analyse für das unbekannte Wort DUNKELBRAUNES als Lösung SUB anbietet; 110/2 gleich zwei dieser falschen SUB-Lösungen – beidemal sind es unbekannte Wörter (HELLGRUENEN bzw. DUNKELGRUENEN) zeigt der sehr lehrreiche Satz

164 ZWISCHEN EINEM GRAUEN ZEBRA UND EINEM HELLGRUENEN BESTEHT EIN KLEINERER UNTERSCHIED ALS ZWISCHEN EINEM BRAUNEN UND EINEM DUNKELGRUENEN.

Sicherlich ließen sich durch differenziertere Analyseregeln einige dieser Fehler reduzieren, etwa auf der Basis der noch unvollkommen erstellten Anreihungsregeln oder der unvollkommen angewendeten Kongruenzregeln. Die bisherige Syntaxanalyse ist noch zu wenig sprachlich-grammatisch adäquat, als daß sie bessere Resultate erlauben würde. Wie schwach oder fehlerhaft die grammatische Reduktionskapazität der Analyse (vielfach auch noch bedingt durch die mit der technischen Komplexität verbundenen Probleme bei der Herstellung eines fehlerfrei programmierten Systems) ist, zeigen Sätze, die augenscheinlich außerhalb der Inventarkontrolle geraten sind, wie

120 SIE WACHSEN DANN FELSENFEST ZUSAMMEN.,

wo das unbekannte Wort FELSENFEST als VRB klassifiziert <S. 111> wurde, obgleich im Satz bereits im ersten Durchlauf WACHSEN als VRB bestimmt worden war. Ein ähnlicher krasser Mißgriff liegt vor in Satz 147, wo ENTVOELKERN als SUB klassifiziert wird, obgleich das unmittelbar davorstehende Wörtchen ZU bereits als 'ZU zum Infinitiv' (IZU mit ZZI als Sonderinformation) bestimmt worden war. 111/1

Dieser kurze, notwendig unvollständige Überblick über die Fehlerursachen mag ausreichen. Der größte Teil ist auf die unvollständige Ausarbeitung der Regeln und Programme, also die Vorläufigkeit der Analyse, zurückzuführen; es ist nicht sehr problematisch, hier, Verfeinerungen oder Verbesserungen soweit vorzunehmen, daß die Ergebnisse deutlich besser ausfallen: Die wirkliche Schwäche liegt in den formal richtigen, sprachwirklich jedoch mehrdeutigen und damit möglicherweise falschen Lösungen begründet. Für diesen Fall sind entsprechende Verfahren zu entwickeln, sei es, daß durch weitere (differenzierende) Informationen bis hin zu semantischen Merkmalen diese formal mehrdeutiggn Fälle eingeschränkt werden, oder sei es, diß ein interaktives Kommunikations-System Mensch-Maschine diese Lücke zu schließen sucht, dergestalt, daß dann der Mensch die noch verbleibenden Mehrdeutigkeiten aus seinem umfachreicheren Wissen heraus entscheidet.2 Ein Ausweg im Hinblick auf eine völlige automatische Klassifizierung bietet sich nur darin, daß die Mehrdeutigkeit – sofern sie wortgebunden ist – im Zweifelsfall entsprechend gekennzeichnet und durch das mögliche wiederholte Auftreten Vereindeutigungen ermöglicht werden. <S. 112>

4.3.3.3 Zum Vergleich: Ergebnisse bei der Analyse von Zeitungstexten

Um das Kurzzeitlexikon erproben zu können, war es nötig, das Textmaterial wenigstens im Hinblick auf die Wortwahl einzugrenzen. Daher konnte man nicht sicher sein, ob die zudem eigens für diesen Zweck konstruierten Testsätze einigermaßen zuverlässige Aussagen über die Brauchbarkeit bei der automatischen Klassifizierung beliebiger Sätze zuließen. Ein Test mittels eines kleinen Corpus von

fortlaufenden Sätzen aus Zeitungs- und Zeitschriftenartikeln soll das Analyseergebnisse ergänzen und bewerten helfen.

Um unterschiedliche sachliche Bereiche berücksichtigen zu können, wurden eine Lokalzeitung 112/1, eine Frauenzeitung 112/2 und eine überregionale Zeitschrift 112/3 zugrundegelegt. Zunächst eine Liste der Artikel mit Quellenangabe und Hinweis auf die entsprechenden Satznummern 112/4 der Analyse:

Nr	*Satz*	*Titel*	*Quelle*	*(Seite)*
1	500-505	Als-ob-Tarife sollen an steigen	SZ	1
2	506-509	USA weiterhin ein Land ohne Personalausweis	SZ	3
3	510-521	Deutsch ist Trumpf in Afrika	SZ	3
4	522-534	In weitem Umkreis klirrten die Scheiben	SZ	8
5	535-544	(Test: Entsafter)	FÜR SIE	76
6	545-550	Steuerflucht	SPIEGEL	60
7	551-556	(Anzeige)	SPIEGEL	20

<S. 113>
Das daraus erstellte Corpus 113/1 umfaßte 57 Sätze mit insgesamt 1155 laufenden Wortformen, hat also ungefähr ein Drittel des Umfangs der Testsätze mit vorgegebenen Wörtern. Augenfällig ist die durchschnittliche Satzlänge von 20,2 Wörtern je Satz; die Zeitungssätze sind also im Mittel fast doppelt so lang wie die Testsätze (11,2). Der relative Anteil der unbekannten, also nicht im Lexikon verzeichneten Wortformen weicht dagegen nur unbedeutend von dem des ersten Testcorpus ab: 16,2 aller Wortformen (gegenüber 17,7 %) waren unbekannt; auch die Anzahl der morphologisch klassifizierten Wörter 113/2 hat sich zwar erhöht (6,2 % des Gesamttextes), bleibt bei einem Anteil von etwa einem Drittel aber noch in der Nähe des bei den Testsätzen ermittelten Wertes (ein Viertel)--für nicht-vorgegebene Wörter. 113/3

Nach allen bisherigen Erfahrungen ließ vor allem die größere Satzlänge auf eine komplexere Struktur und damit eine Verringerung des Anteils richtiger Lösungen schließen. Hinzu kommt, daß die Zeitungssätze, einen durchschnittlichen Anteil von 3,2 unbekannten Wörtern je Satz aufwiesen, während auf die Testsätze im Mittel zwei unbekannte Wörter je Satz kamen. Selbst wenn die morphologisch eindeutig klassifizierbaren Wörter ausgeklammert werden, ist der Anteil von zwei durch die Analyse zu klassifizierenden unbekannten Wörtern (gegenüber 1,7 bei den Testsätzen) noch erhöht.

Ehe auf Einzelheiten eingegangen wird, soll eine Übersicht die wichtigsten Analyseergebnisse 113/4 aufzeigen: <S. 114>

HO-TYP	*D*	*Belege*	*richtig*	*rel.*	*Test 1*	*Syntaxanalyse*
5	II	3	3	100 %	90 %	93 %
10	V	21	17	80,4%	80 %	77 %
11	III	17	14	82,2%	91 %	91 %
14	V	57	41	71,9%	73 %	74 %
46	III	11	11	100 %	77 %	87 %
55	II	6	5	83,3%	-	-
HO ges.		115	91	79,1%	80 %	88 %

Morph. A.	72	60	84,4%	90 %
Ges.	187	151	80,8%	

Obwohl bei der geringen absoluten Zahl einiger Einzelwerte wenig sichere Grundlagen zum Vergleich gegeben sind, lassen sich einige interessante Phänomene aufzeigen:

Die durchschnittliche Lösungsquote der Testsätze wird nahezu erreicht. Legt man die Erfahrung zugrunde, daß längere Sätze normalerweise komplexere Strukturen aufweisen als kürzere Sätze, so wird deutlich, daß die durchschnittlich kürzeren, eigens zur Analyse konstruierten Testsätze durchaus mit dem Schwierigkeitsgrad dieser beliebigen Sätze vergleichbar sind.

Bereits für die morphologische Analyse läßt sich zeigen, daß das Ergebnis von 84 % richtiger Lösungen nur als untere Grenze zu werten ist und vorwiegend dem Modellcharakter des Verfahrens Rechnung trägt: Bei einer feineren morphologischen Klassifizierung hätten EINGRENZEN (529, ENZ=SUB), ENTSENDEN (511, END=ADJ), FRUCHTFLEISCH (539, ISCH=ADJ), SAARWIRTSCHAFT (504, HAFT=ADJ), SIEBKORB (540, 542, SIEB als morphologisches Kriterium für die Zahl SIEBEN ...), SONNTAGABEND (529, END=ADJ) und STEUEROASENGESETZ (550, Z-NAM) nicht vorzeitig fälschlich bestimmt zu werden brauchen; auch die theoretische Voraussetzung des Modells, <S. 115> daß alle Partizipien starker Verben lexikalisch erfaßt seien, führt – wie schon bei den Testsätzen – notwendig zu Fehlern: bei AUFGEHOBEN (544, EN≠PTZ) und VORAUSGESEHEN (517, EN≠PTZ). Bereits bei Vermeiden dieser Fehler durch eine exaktere morphologische Präanalyse hätte die Fehlerquote des Gesamtergebnisses um 1/4 gesenkt werden können.

Einfache Möglichkeiten, die Fehlerquote weiter herabzusetzen, zeigen sich, wenn man die Liste der nach Wortklassen aufgegliederten Wortformen betrachtet:

Wortklasse	*Belege*	*davon falsch*
SUB	91	15
ADJ	11	4
VRB	4	2
ADV	1	1
PTZ	7	2
INF	1	-

Besonders auffallend ist der hohe Anteil der Substantive an unbekannten Wörtern. Während er bei den Testsätzen bei nahezu 50 % lag, beträgt er bei den Zeitungstexten 80 %. 115/1 Das Adverb, mit etwa 10 % Anteil immerhin noch an vierter Stelle gelegen, kommt nur noch einmal (VORWEG, 535, fälschlich als SUB gelöst) vor; nur vier (!) der 115 unbekannten Wortformen sind finite Verben. 115/2 Wenn beim Aufbau des Lexikons und der Satzanalyseprogramme diesem Phänomen Rechnung getragen würde, ließen sich auch hierdurch die Analyseergebnisse noch verbessern: Immerhin wurden sieben Substantive (von 15) fälschlich <S. 116> als Adverbien klassifiziert. Zumindest ließe sich durch eine vollständigere Aufnahme der Adverbien ins Lexikon die Mehrdeutigkeit unbekannter Wortformen verringern.
Im wesentlichen bestätigen sich darüberhinaus die Ergebnisse der Analyse der Testsätze: So wurden alle möglichen PTZ erkannt, 116/1 auch die Lösungsquote der

Substantive (hier 83,5 %) liegt nur etwas über derjenigen für die Testsätze. Die geringeren Belegzahlen für ADJ, VRB, INF und ADV lassen dagegen kaum Vergleiche zu. 116/2

Voraussetzung für eine (hier nicht in letzter Konsequenz angestrebte, sondern nur aufgezeigte) Verbesserung der Analyseresultate sind eine den theoretischen Anforderungen an das Basislexikon stärker entsprechende Lexikonstruktur, ein feineres morphologisches Analysesystem auf der Basis eines Stamm- oder Morphenwörterbuchs (das vor allem Erkennungsroutinen für Nominalkomposita enthalten müßte) und nicht zuletzt eine exaktere, evtl. den speziellen Bedingungen der dann noch nicht erfaßbaren Wörter angepaßte Analyse. Da das gegenwärtig funktionsfähige Saarbrücker Analysemodell diese Voraussetzungen nicht in hinreichendem Maße aufweist, waren weitgehend fehlerfreie Ergebnisse nicht zu erwarten. Die erreichten Werte ermutigen aber dazu, den bisher eingeschlagenen Weg fortzusetzen.

4.3.3.4 Ergebnisse bei der Verwendung des Kurzzeitlexikons

Im folgenden soll auf die Veränderungen des Kurzzeitlexikons während der elf Analysezyklen des zweiten Gesamtlaufs der Testsätze und vor allem auf den endgültigen Stand des Lexikons eingegangen werden. Die Ergebnisse für die nicht vorgegebenen unbekannten Wörter, die ebenfalls automatisch in das Kurzzeitlexikon übernommen wurden, sind hier zu <S, 117> vernachlässigen. Wie zu erwarten war, erreichte keines dieser unbekannten Wörter eine fortgeschrittene Phase: Nur neun der insgesamt 206 nicht vorgegebenen unbekannten Wörter (CAFE, COMPUTER, FLECHTEN, KOSTUEM, SAEGE, SCHECK, TATAREN, VOLKSWAGEN, WERKZEUGE) waren zweimal, der Rest, also 197, nur einmal belegt; in diesen Fällen spiegelt das Analyseergebnis also zugleich den Stand des Kurzzeitlexikons wieder. Einzelheiten zu den entsprechenden Analyseergebnissen sind in Tabelle V aufgeführt.

Alle 26 vorgegebenen unbekannten Wörter waren dagegen mehr als achtmal (die erforderliche Mindestzahl für das Erreichen der Endphase) belegt, die Mindestbelegzahl war 10 (BUNDESPOLITIKER 117/1), alle übrigen waren in einem Zyklus mindestens einmal vertreten, keine Wortform war jedoch mehr als zwölfmal insgesamt belegt.

Bevor jedoch die Ergebnisse des Kurzzeitlexikons näher betrachtet werden sollen, sei noch einmal auf die diesbezüglichen Analyseergebnisse ohne lexikonverwendung zurückgegangen.2 Für vier der vorgegebenen Wörter war von Vornherein – dies zeigt bereits die Analyse – kein Erreichen der Endphase, in der noch nicht belegte Wortklassen 'verboten' werden, zu erwarten, da der erforderliche Mindestwert (Schwellenwert) aufgrund der hohen Fehlerzahl nicht erreicht wurde. Es sind die Wortformen BEDECKEN (VRB/INF) (7 richtige von 11 möglichen richtigen Lösungen), DUNKELBRAUN (ADV) (5 von 11), FELSENFEST (ADV) (3 von 11) und KOSAKEN (SUB) (4 von 11). Alle gehörten sie zu den fünfdeutigen Homographen; die Schwierigkeit, ein ADV zu erkennen, wird bei DUNKELBRAUN und FELSENFEST wiederum offenbar. Darüberhinaus reichten die neun richtigen Lösungen bei dem auch 'natürlich mehrdeutigen' Wort ANLAECHELN nicht aus, da sich fünf auf die Klasse VRB, <S. 118> vier auf die Klasse SUB (Substantivierter Infinitiv) bezogen. 118/1

Von vornherein stand dagegen zu erwarten, daß die Wörter ABDRUCK (HO 11 – SUB), GRINST (HO 10 – VRB), PHONEM (HO 5-em – SUB) sowie VERBEUGE (HO 14-e – VRB) die Endphase erreichen konnten, da alle Lösungen eindeutig und richtig waren; zweifelhaft war hier nur, ob genügend sichere Lösungen vorlagen. 'Fehlerfrei war auch noch die Wortform DUNKELBRAUNE (14-e) gelöst worden, wobei allerdings den neun ADJ-Belegen zwei SUB-Belege gegenüberstanden. In allen übrigen (17) Fällen war es zweifelhaft, ob eine eindeutige Reduktion durchgeführt werden könnte, da mindestens ein Fehler vorgekommen war.2

Zunächst eine tabellarische Übersicht über die Wortformen, bei denen im Kurzzeitlexikon für eine Wortklassenlösung eine Endphase erreicht wurde:

Wortform	*alter HO-Typ*	*im Verlauf ausgeschl.*	*noch erlaubte WK 118/3*	*neuer HO-Typ*
ABDRUCK	11	ADV	SUB	-
ANALOGER	46	ADV	*ADJ*/SUB	5
ANLAUFS	11	ADV	SUB	–
AXT	10	VRB	ADV/PTZ/SUB	9
BERGAB	11	SUB	ADV	-
BUNDESPOLITIKER	46	-	ADJ/ADV/SUB	46
DUNKELBRAUNE	14-e	ADV/VRB	ADJ SUB	5
DUNKELBRAUNEM	5-em	-	ADJ SUB	5
DUNKELGRUENEN	14-en	ADV/VRB/INF	ADJ/SUB	5
EINGELOEST	10	(ADV)/VRB4	PTZ/SUB	9 <S. 119>

Wortform	*alter HO-Typ*	*im Verlauf ausgeschl.*	*noch erlaubte WK*	*neuer HO-Typ*
GRINST	10	ADV/PTZ/SUB	VRB	-
KRAXELT	10	ADV/PTZ/SUB	VRB	-
MACHTKOMPLEXES	5-es	-	ADV/*SUB*	5
PEST	10	ADV/PTZ	VRB/*SUB*	2
PHONEM	5-em	ADJ	SUB	-
VERBEUGE	14-e	ADV/ADJ/SUB	VRB	-

Für sieben dieser 16 Wortformen haben also die Belege ausgereicht, um ein Wort endgültig und eindeutig nach der Wortklasse automatisch zu bestimmen. Dies war nur dadurch möglich, daß acht sichere Lösungen für diese eine Wortklasse gefunden waren (einbegriffen : zwei unsichere = eine sichere Lösung), ehe womöglich ein Fehler auftrat. Bei ABDRUCK, GRINST, PHONEM und VERBEUGE war, wie schon die Analyse 'ohne Kurzzeitlexikon' annehmen ließ, kein Fehler aufgetreten. Die in der Analyse 'ohne Kurzzeitlexikon' aufgetretenen Fehler ANLAUFS --- ADV (Satz 236); BERGAB --- SUB (Sätze 232, 236) und KRAXELT --- PTZ (Satz 227) blieben beim Einbeziehen des Lexikons ohne Einfluß, da vorher schon für diese Wörter die Endphase erreicht und die entsprechende Klasse verboten worden war. So wurden diese Fehler auch bei der Analyse einschließlich des Kurzzeitlexikons nicht mehr begangen: Diese Sätze sind also im zweiten Gesamtdurchlauf der elf Zyklen richtig gelöst worden.

Wenn auch nicht alle Wörter eindeutig klassifiziert werden konnten, so sind andererseits auch keine 'richtigen' Lösungen fälschlich ausgeschlossen worden.

Hier soll kurz auf einen Sonderfall der Reduktion eingegangen werden: Prinzipiell werden beim Eneichen einer Endphase für eine Wortklasse alle anderen Wortklassen – sofern sie nicht schon über die quasimorphologische <S. 120> Analyse verboten worden waren – ausgeschlossen, wenn sie bis dahin noch nicht 'sicher' belegt sind. Dennoch sollte der generellen Mehrdeutigkeit auch in diesem Verfahren Rechnung getragen werden. Endet beispielsweise ein VRB auf EN, so sind – selbst wenn nicht belegt – der Infinitiv und die Substantivierung zuzulassen. Ähnliches gilt für PTZ, das nicht nur als Teil der Verbalgruppe, sondern auch als ADV vorkommen kann. Daher wurden diese generellen Mehrdeutigkeiten trotz eines vorher möglichen Verbots bei der Wörterbuchzuordnung wieder für die Analyse zugelassen. 120/1

In sechs Fällen wurden bei Erreichen der Endphase zu einem Wort andere – noch nicht belegte – Wortklassen verboten, d.h. die Deutigkeiten konnten für diese Wortformen zumindest reduziert werden. Dies geschah für die Wörter ANALOGER, AXT, DUNKELBRAUNE, DUNKELGRUENEN, EINGELOEST und PEST. Diese Einengung der Mehrdeutigkeiten bringt es nach den bisherigen Erfahrungen mit sich, daß die Fehleranfälligkeit sinkt, also die Analyseergebnisse gleichfalls verbessert werden können.

Betrachtet man die insgesamt neun noch mehrdeutigen reduzierten Belege näher, so ist auffällig, daß fünf davon die Mehrdeutigkeit ADJ/SUB aufweisen. Bei dreien (DUNKELBRAUNE, DUNKELBRAUNEM, DUNKELGRUENEN) ist diese verbliebene Mehrdeutigkeit eine Folge der möglichen Substantivierung der Adjektive; bei diesen dreien ist zugleich die Endphase für das Adjektiv erreicht; ähnliches gilt für ANALOGER. Beim fünften Element dieser Gruppe, MACHTKOMPLEXES, ist die Endphase für das richtige SUB erreicht. Klammert man einmal die generell mögliche Substantivierbarkeit der Adjektive aus, so wird bei der näheren Betrachtung dieser neun nicht eindeutig reduzierten offenkundig, daß bei allen die Wortklasse, die in die Endphase getreten <S. 121> ist., zugleich diejenige ist, die als endgültig reduzierte hätte erscheinen müssen. Die übrigen beruhen auf falschen Lösungen. Doch es ist noch ein interessantes Phänomen zu beobachten: Keine der reduzierten Wortformen weist mehr als einen Falschbeleg für eine andere Wortklasse auf. Wäre daher also eine entsprechende Regel vorhanden gewesen der Art, daß selbst dann eine Wortklasse bei acht Belegen für eine andere Wortklasse verboten werden dürfte, wenn ein sicherer Beleg vorhanden ist, dann wären bis auf DUNKELBRAUNE, das zweimal SUB (Substantivierung) aufweist, alle reduzierten Wortklassen eindeutig und richtig bestimmt gewesen.

Diese 'Regel' hätte also einen bedeutend größeren Effekt erzielt als die Unterscheidung und Berücksichtigung 'sicherer' und 'unsicherer' Lösungsergebnisse, wie sie für den Test vorgenommen war. In keinem der Fälle hat mittels dieses Merkmals ein Beleg ausgeschlossen werddn können, da keine einzige dieser 'Falschlösungen' nur als 'unsicher' gekennzeichnet werden konnte, wie überhaupt dieser Zweifelsfall sehr selten auftrat. 121/1 Alle falschen Lösungen waren nach der Analyse also hier vermeintlich 'sicher', während ein geringer Teil der richtigen Lösungen als 'unsicher' klassifiziert wurde, was hin und wieder den Reduktionsprozeß verzögerte.

Dies stellt m.E. diese Methode nicht prinzipiell in Frage, sondern zeigt eher, daß sie für das gegenwärtige Saarbrücker Modell in dieser Form unbefriedigend ist, solange nicht die 'sicheren' Lösungen eingeschränkt werddn können. In jedem Falle scheint es nötig zu sein, hier neben rein linguistischen auch statistische (empirische)

Methoden zu erproben, etwa der Art, wie sie oben für die 'erwünschte' Regel verwendet worden sind. <S. 122>

Bei den zehn verbleibenden Wortformen, die nach den elf Zyklen (noch) nicht zu einer Mehrdeutigkeitsreduktion geführt haben, hält es schwer, irgendwelche Tendenzen oder Möglichkeiten aufzuzeigen. Dies zeigt eine Übersicht der Lösungen bzw. der möglichen richtigen Ergebnisse:[122/1]

Wortform	*HO*	*ADJ*	*ADV*	*VRB*	*SUB*	*PTZ*	*INF*
ANGSTHASE	14	- (-)	3 (-)	2 (-)	7 (12)	*	*
ANLAECHELN	14	*	1 (-)	6[122/2] (-)	5 (4)	*	- (2)
BEANSTANDETEN	14	5 (6)	- (-)	7 (6)1	1 (-)	*	- (-)
BEDECKEN	14	1 (-)	- (-)	7 (9)	3[122/2] (-)	*	1(2)
BEGAFFT	10	*	2 (-)	7[122/2] (9)	1 (-)	1 (2)	*
DUNKELBRAUN	14	1[122/3] (*)	4 (7)	3 (-)	3 (4)	*	- (-)
DUNKELBRAUNES	5	7 (9)	*	*	4 (2)	*	*
FELSENFEST	10	*	4 (10)	2 (-)	4 (1)	1[122/2] (-)	*
GELBER	46	7 (10)	2 (2)	*	3 (-)	*	*
KOSAKEN	14	1 (-)	1 (-)	5 (-)	4[122/2](11)	*	- (-)

Trotz dieser verwirrenden und statistisch wenig ergiebigen Zahlen sollen einige Betrachtungen angestellt und ein paar Vergleiche gezogen werden: Am ehesten wäre noch bei BEANSTANLETEN und ANLAECHELN ein relativ gutes Ergebnis festzustellen; nur einmal ist eine sprachlich nicht mögliche Wortklasse bei BEANSTANDETEN (SUB) aufgetreten. Hier verhindert die natürliche Mehrdeutigkeit dieses Wortes (ADJ/VRB) eine vorzeitige Reduktion; ähnliches gilt für ANLAECHELN, das häufig in substantivischem Gebrauch (Substantivierung) belegt ist. Bei einer einzigen weiteren sicheren Lösung VRB würde bereits bei BEANSTANDETEN das ADV ausgeschlossen, während der hier nicht mögliche (und auch nicht <S. 123> fälschlich bestimmte) INF vorläufig als 'generelle' Möglichkeit noch zugelassen wäre, da die Entscheidung VRB in Verbindung mit der Endung -EN diese Möglichkeit noch vorsieht. Es wäre einmal empirisch zu prüfen, ob und wann eine fehlende Infinitivlösung darauf schließen ließe, daß – wie hier – eine Imperfektform vorliegt, eine wichtige Information für eine etwaige automatische Reduktion auf die Grundform.

Drei weitere Wortformen (ANGSTHASE, BEDECKEN und DUNKELBRAUNES) zeigen immerhin noch Möglichkeiten, eine Reduktion über die absolute Lösungshäufigkeit auf empirischer Basis zu entwickeln, da die häufigste Lösung zugleich mit der richtigen übereinstimmt. Bei den natürlich mehrdeutigen Wortformen BEGAFFT (VRB/PTZ) und GELBER (ADJ/ADV) zeigt sich jedoch bereits die Grenze eines derartigen Verfahrens, da damit die Gefahr besteht, daß sprachlich mögliche Deutigkeiten, also natürliche Mehrdeutigkeiten, ausgeschlossen werden könnten.

Die drei noch verbleibenden Wortformen FELSENFEST, DUNKELBRAUN und KOSAKEN fallen völlig aus dem Rahmen. Sieht man einmal von dem metasprachlichen Gebrauch in

42 BILDEN SIE NOCH EINEN SATZ MIT ' FELSENFEST '!

(falsche Lösung: SUB) ab, so bleiben bei FELSENFEST immer noch 6 falsche Ergebnisse übrig. Hier wie bei DUNKELBRAUN – das allerdings als Farb-ADV noch zusätzlich (mögliche und häufige Substantivierung) generell mehrdeutig zu sein scheint – zeigt sich die Schwäche der Analyse, Adverbien richtig zu erkennen. KOSAKEN, das einzige Substantiv unter den fünfdeutigen künstlichen Homograpien vom Typ HO 14, könnte möglicherweise aufgrund einer bei Substantiven sonst unüblicheren Stellung in der Wortfolge bei der Analyse zu dieser hohen Fehlerzahl geführt haben: In 6 (!) der 7 falschen Lösungen steht kein Artikel, Indefinitpronomen oder Adjektiv vor diesem Wort, während bei allen <S. 124> vier richtigen Lösungen eine derartige Wortklasse vorausgeht. Hier wäre etwa ein Anzeichen dafür zu sehen, daß die überfließenden Deklinations- und Konjugationsangaben der 'unbekannten' Wörter wegen ihrer generellen Kongruenzfähigkeit die Fehlerquote erhöhen können. Einschränkungen sind hier durchaus möglich; so wäre in vielen Fällen der Plural für ein mögliches Substantiv (AXT, BERGAB,...) auszuschließen, eine Differenzierung, wie sie für das beschriebene Modell. nicht vorgenommen worden ist.

Eine letzte Übersicht soll noch einmal den Zusammenhang und die Kontraste in den einzelnen Reduktionsergebnissen verdeutlichen:

HO-Typ	*Deutigk.*	*Belege insges.*	*ganz reduz.*	*teilw. redz.*	*nicht reduz.*
5	II	4	1	2	1
10	V	7	2	3	2
11	II	3	3	-	-
14	V	9	1	2	6
46	III	3	-	2	1
GES.	*26*	*7*	*9*	*10*	

Obwohl die geringe Zahl der Belege natürlich keine sicheren Schlüsse zuläßt, dürften einige Tendenzen und Erfahrungen noch einmal deutlich werden: Während die sieben zwei-deutigen Homographen bis auf das Wort DUNKELBRAUN (Substantivierung!) zumindest in der Mehrdeutigkeit reduziert wurden (die drei Belege für HO 11 werden sogar alle eindeutig klassifiziert), ist gerade der fünfdeutige Homograph vom Typ HO 14 in sechs Fällen nicht reduzierbar und nur in einem Fall eindeutig bestimmbar gewesen. Dies zeigt also noch einmal, daß die Anzahl der Lösungsmöglichkeiten die Fehlerhäufigkeit ansteigen läßt. <S. 125>

Daß die Lösungsmöglichkeit ADV bei HO 46 nicht die Hauptursache für eine fehlende endgültige Klassifikation einer der entsprechend 'künstliche mehrdeutigen Wortformen war, zeigen die Wortformen ANALOGER, wo kein Beleg für ADV fälschlich auftrat, und GELBER, wo die zwei Belege für ADV der natürlichen Lösungsmöglichkeit (und Mehrdeutigkeit) dieses Wortes entsprechen. Nur einmal (AXT in Satz 143) wurde bei allen Homographen vom Typ HO 46 fälschlich auf ADV entschieden. Mit Ausnahme von HO 14 konnten alle übrigen künstlichen Mehrdeutigkeiten zu zwei Dritteln bereits durch die 11 bzw. 12 Belege anhand des Kontexts in ihrer Deutigkeit reduziert werden, ohne daß dabei eine 'richtige' Lösung fälschlich ausgeschlossen worden wäre. Wenn auch noch eine Reihe von Fragen offenbleiben, Wenn auch vielleicht eine Anzahl von Möglichkeiten bleibt, dieses Verfahren zu modifizieren, so, scheinen doch der Analysetest und auch die Ergebnisse des Kurzzeitlexikons zu zeigen, daß sich für die maschinelle

Sprachanalyse bei einem derartigen Lexikonverständnis neue Möglichkeiten ergeben.

4.4 Mensch und Computer

Mit dem Modell zu einet' automatischen Klassifikation unbekannter Wörter ist zu zeigen versucht worden, daß es – zumindest im Bereich der Syntax – möglich erscheint, die Lücke zwischen einem konkreten und einem idealen Lexikon in einigen Fällen der maschinellen Sprachbearbeitung zu überbrücken. Es scheint, daß damit das (deskriptive) Problem der TG, neben einem möglichst vollständigen Regelsystem auch ein möglichst kgmplettes Lexikon bereitzustellen, das etwa auch die zwar sprachlich noch nicht realisierten, aber möglichen Einträge umfaßt, in manchen Fällen als ein explanatives, der Spracherlernung zugehöriges Problem zu erklären ist. Natürlich wurde hier kein Sprach- oder besser: Wortschatzerkennungssystem der Art entwickelt oder vorgestellt, wie es etwa ein Kind zum Aufbau seiner <S. 125> Sprache benötigt. Ein derartiges System für einen Computer in vollem Umfang zu entwickeln, hieße seine derzeitigen Möglichkeiten wohl überschätzen, wenn auch die heutigen technischen Gegebenheiten für ein Dialogsystem Mensch-Maschine erste Voraussetzungen dazu geschaffen haben. Das hier vorgelegte Modell könnte – wollte man es auf die menschliche Fähigkeit des Wortschatzerwerbs beziehen – eher mit den Strategien eines Erwachsenen verglichen werden, der beim lesen eines Buches – etwa in einer Fremdsprache – auf ein ihm noch unbekanntes (oder auch wieder völlig vergessenes, also nicht zu seinerä Basiswortschatz gehörendes) Wort trifft und es anhand des Kontexts syntaktisch (allerdings im Gegensatz zu dem Computermodell auch semantisch) zu klassifizieren, das heißt in ein bestehendes grammatisches System als ein Element einer bekannten Gruppe einzuordnen versucht.

Dem hier vorgelegten Modell fehlen natürlich noch einige wesentliche Grundlagen, um hier auch einem echten Vergleich Mensch-Maschine standzuhalten. Die erworbenen Inform<tionen sind noch sehr dürftig, wie auch die Mittel der morphologischen und syntaktischen Analyse noch recht bescheiden sind. Doch scheint es, als stünden wir bei dieser Art der Informationserschließung, also des automatischen Aufbaus eines maschinellen Lexikons, erst am Anfang. Wie Quillians Versuch, einen Computer mithilfe eines umfangreichen sprachlichen Relationensystems etwas Verständnis für semantische Zusammenhänge einzugeben, 126/1 gezeigt hat, steht neben einer Reihe von Möglichkeiten auch noch eine Anzahl von Problemen offen, die nicht nur technischer Natur (etwa begrenzte Speicherkapazität der Maschine), sondern linguistischer Art (fehlende Wohldefiniertheit sprachlicher und metasprachlicher Termini, zu oberflächliche oder unzureichende <S. 127> Merkmalbeschreibungen) sind. Es scheint auch, daß die Verwendung des Computers in der Lexikologie 127/l zu systematischeren, den modernen Sprachtheorien angemesseneren humanen 127/2 Lexika führen kann. Der Computer bietet, wie es schon Lenders richtig eingeschätzt hat, 127/3 die Möglichkeit, von einem unbeweglichen, starren Lexikon wegzukommen und zu einem beweglicheren, flexiblen Lexikonsystem zu gelangen, das den Wandlungen des Wortschatzes oder den textspezifischen Gegebenheiten besser gerecht wird.

Dazu einen Beitrag zu leisten, war das Ziel dieser Arbeit.

Fußnoten:

IV/1 Die in Klammern einegschlossene Angabe bezieht sich auf die Zitierweise in der Arbeit.

1/1) In der Folge wird eine gewisse Vertrautheit mit dem Aufbau und der Arbeitsweise des Computers sowie seiner Funktion in der Linguistischen Datenverarbeitung vorausgesetzt. Ausführliche Einführungen dazu geben u.a. LAMB (Digital Computer) und HAYS (Computational Linguistics). An dutschsprachigen ist vor allem zu nennen DWORATSCHEK (Datenverarbeitung).

1/2) Bei den ersten, z.T. großangelegten Projekten (etwa dem Georgetown-Projekt) handelte es sich in der Regel um 'rohe' Wort-für-Wort-Übersetzungen. Zur Geschichte der MT vgl. SCHNELLE (Theorie). Zur Entwicklung bis ca. 1960 s.a. BAR-HILLEL (Translation) und PENDERGRAFT (Languages).

1/3) Deutsche Kurzfassung s. SPRACHE UND MASCHINEN.

2/1) Die gegenwärtig noch laufenden Vorhaben - vielleicht abgesehen von der Toma-Version des Georgetown-Projekts in den USA, der in Ispra (Italien) verwendeten MT und des bei der IBM in der Patentdokumentation benutzten Übersetzungsverfahrens Englisch-Deutsch - dienen m.W. vorwiegend der Grundlagenforschung (etwa das Saarbrücker Projekt Russisch-Deutsch der Erprobung eines Transfer- und Synthese-Modells). Vgl. TOMA (SERNA-System) und JANDA (Dokumentation); PERSCHKE (Sprachübersetzung); SCHIRMER (Übersetzungssystem).

2/2) Zur Vielzahl der Programmiersprachen vgl. ausführlich SAMMET (Programming Languages). Die Frage der Verwendbarkeit einzelner Programmiersprachen in der LDV ist behandelt von RASKIN (Humanities).

3/1) Ich stütze mich im wesentlichen auf die Theorie, wie sie CHOMSKY (Aspekte) 1965 formulierte. Die seitherige Entwicklung spielt für die spezifischen Fragestellungen dieser Arbeit nur eine geringe Rolle - mit Ausnahme der verschiedenen Arbeiten zur Struktur des Lexikons wie BOTHA (Lexicon) und GRUBER (Relations), auf die von Fall zu Fall eingegangen wird.

4/1) Ein Computer dieser Bezeichnung stand der Saarbrücker Arbeitsgruppe für die maschinelle syntktische Analyse zur Verfügung. Vgl. auch Kap. 4.1.1 S. 70 Anm. 1.

4/2) Einen ausführlichen Bericht über Ergebnisse und Vorgehensweise dieses Unternehmens, das zu Anfang der sechziger Jahre begonnen wurde, geben EGGERS et.al. (Syntaxanalyse).

7/1) CHOMSKY (Aspekte) S. 15.

7/2) Vgl. zum folgenden ausführlich CHOMSKY (Aspekte), Kap. 1,1,

7/3) CHOMSKY (Aspekte) S. 181. Feststellungen dieser Art sind jedoch schon älter. Vgl. den Hinweis Chomskys auf Bloomfield a.a.O.S. 267.

8/1) + bedeutet 'positiv spezifiziert in bezug auf das betreffende Merkmal ; - bedeutet 'negativ spezifiziert'.

8/2) FRIEDMAN (Application) S. 15, Teil von Fig.9. Ich wähle zur Veranschaulichung allerdings ein Stemma.

8/3) FRIEDMAN (Application) S. 6, Fig. 3.

9/1) Die kontextsensitiven Merkmale [-(SN(DET))] können in diesem Zusammenhang vernachlässigt werden.
9/2) CHOMSKY (Aspekte) S. 181.
9/3) CHOMSKY (Aspekte) S. 114 und 181.

10/1) CHOMSKY (Aspekte) S. 113.

11/1) Vgl. unten Kap. 2.1.3.
11/2) Ganz so einfach.. ist dies natürlich nicht; so muß etwa ein Merkmal PLURALFÄHIGKEIT implizit oder explizit vorhanden sein (Es gibt Nomen, die keinen Plural bilden können, z.B. FURCHT), während eine Einschränkung der Kasusfähigkeit im Deutschen nicht existiert.
11/3) Vgl. dazu ausführlich CHOMSKY (Aspekte) S. 188 ff.

12/1) Eine andere Relation von SELBST' (nicht zur Charakterisierung des Verbs, sondern des Objekts) im Sinne von 'PERSOENLICH' ist allerdings noch möglich. Etwa: ER HOERT MICH (SELBST) UND NICHT EIN VON MIR BESPROCHENES TONBAND.
12/2) Vgl. dazu ausführlicher und auf die Problematik eingehend KLEIN (Studien) S. 119-132 (=Kap. 4.4).

13/1) CHOMSKY (Aspekte) S. 206 ff.
13/2) Bei Ausnahmen wie Leichnam, Toter müssen evtl. beide Merkmale, alsot [+menschlich], [-belebt], angegeben werden.

14/1) CHOMSKY (Aspekte) S. 212.
14/2) Wir wollen den hier naheliegenden Gedanken nicht weiter verfolgen, inwieweit diesem Redundanzprinzip Vorgänge im menschlichen Gehirn parallel zu ordnen sind. Vgl. den Hinweis BIERWISCHS darauf, daß Kinder zunächst einfachste (morphol.) Regeln anwenden, etwa die der schwachen Verbflexion (ich LAUFTE statt ich LIEF): BIERWISCH (Strukturralismus) S. 131 ff.
14/3) DIETRICH (Starke Verben) S. 24.

15/1) DIETRICH (Starke Verben) Liste S. 35 ff.
15/2) BILLMEIER (Simulation) S. 104.
15/3) DIETRICH (Starke Verben) S. 32-34.

16/1) In einfachen Phrasenstrukturgrammatik-Modellen werden Lexikon (-Regel) und Grammatikregel gar nicht getrennt; Vgl. etwa V.HELD (Phrasenstrukturgrammatik) S. 143 ff.

17/1) Mit in die Entscheidung einzubeziehen wären etwa Faktoren wie die Zeit, die der Programmierer für das Erstellen des Regel-Programms benötigt oder die Zeit, die der Kodierer für das Ermitteln des Lexikoneintrags braucht.
17/2) DUDEN (Grammatik) S. 347.
18/1) Das Merkmal kann natürlich implizit in anderen enthalten sein, falls sich eine generalisierende Regel findet.
18/2) CHOMSKY (Aspekte) S. 230 ff.
18/3) CHOMSKY (Aspekte) S. 214.

19/1) Vgl. dazu BOTHA (Lexicon), bes. Kap. IV, wo für die Komposita nachzuweisen versucht wird, daß sie zumeist ins Lexikon aufgenommen werden müssen, da sich kaum allgemeine Regeln formulieren lassen, die ihre Funktion aufgrund der beteiligten freien Morpheme erklären würden. Zu diesem Komplex s.a. CHOMSKY (Nominalization) und WUNDERLICH (Nominalisierungen).

20/1) CHOMSKY (Aspekte) S. 233.
20/2) Beispiele für derartige Ambiguitäten finden sich in großer Zahl bei CLYNE (Komposita), wo auch Lösungsmöglichkeiten gesucht werden. Zu den Möglichkeiten der Regularisierung der Komposita s.a. KASTOWSKY (Wortbildung).

21/1) Die ersten MT-Projekte (Georgetown) benutzten bereits 'Stammwörterbucher' mit morphologischen Angaben zur Zuordnung flektierter Wortformen,
21/2) EGGERS (Syntaxanalyse) S. 55 f.
21/3) Vgl. SCHIRMER (Überwetzungssystem) S. 602.

22/1) Dem widerspricht nicht, daß bereits eine Reihe von Anwendungen der TG, Ja Umsetzungen von TG-Grammatiken in Computerprogramme existieren. Ich erinnere hier noch einmal an die von FRIEDMANN in ihrem TG-Testsystem verwendeten Grammatiken von ROSENBAUM, TRAUGOTT, KLEVANSKY u.a. Vgl. FIEDMANN (Application) S. 18 ff.
22/2) Ein Beispiel ist die Grammatik von GROSS (Verbe).
22/3) Ich will hier nicht für oder wider die Argumente BARHILLELS streiten, der eine maschinelle Übersetzung prinzipiell für unmöglich hält. Selbst wenn eine eindeutige Übersetzung im theoretischen Sinn nicht möglich wäre, lassen sich doch - wie es zumindest das in Ispra arbeitende MT-Verfahren zeigt - praktikable Übersetzungen vorstellen. Zu der ITT in Ispra vgl. PERSCHKE (Sprachübersetzung).

23/1) Daß der Computer gerade bei der umfassenden Datensammlung besonders nützlich ist, braucht nicht weiter betont zu werden. Vgl. dazu GARVIN (Computer) der einige wesentliche Verwendungsmöglichkeiten in der deskriptiven Linguistik aufzeigt.

24/1) Vgl. etwa FILLMORES 'Tiefenkasus', FILLMORE (Case); mit Orientierung an deutschen Verben jetzt auch DIETRICH (Studien)
24/2) CHOMSKY (Aspekte S: 34 ff.
24/3) CHOMSKY (Aspekte S. 35.

26/1) Beide Wörterbücher haben verschiedene Zielsetzungen. Der DUDEN begnügt sich vorwiegend mit der Darstellung der korrekten Schreibweise, gelegentlich ergänzt um kurze Bedeutungshinweise, im 'Wahrig' finden sich darüberhinaus Bedeutungserläuterungen mit Kontextbeispielen. Es geht hier nicht darum, beide Lexika einander gegenüberzustellen, sondern beide in der Adresskomponente den Forderungen der TG zu stellen.

27/1) Ähnlich den Verben verfahren beide Lexika auch bei unregelmäßigen Adjektiven. Im Gegensatz zum DUDEN hat der 'Wahrig' einen Eintrag für BESSER mit Hinweis auf, GUT usw.
27/2) DUDEN (Rechtschreibung) S. 11 (Vorwort)..

27/3) Der Test wurde durchgeführt anhand eines lemmatisierten Index zu Georg Trakls Dichtungen, der mir bereits als Computerausdruck vorlag. Der Index erscheint in Kürze in der Reihe Indices zur deutschen Literatur. Vgl. KLEIN (Trakl). Gerade ein relativ esoterischer und ungewöhnlicher Wortschatz schien mir hier Material zu versprechen, um das Problem der 'Lücken' im Lexikon deutlich werden zu lassen.

29/1) Vgl. WAHRIG (Wörterbuch), Vorwort (ohne Seitenangabe).

30/1) Vgl. EGGERS (Syntaxanalyse) S. 55 f.
30/2) Zugrundegelegt werden alle im Wahrig' stehenden (ca. 9o.ooo) Stichwörter mit Ausnahme der dort verzeichneten Abkürzungen, die durch spezielle Verzeichnisse erfaßt werden sollen. Vgl. dazu LEiöMATISIERUNG, wo auch die Behandlung der Starken Verben und anderer Unregelmässigkeiten in der Flexion beschrieben wird.

31/1) ANTAL (Dictionary) scheint einen solchen Idealfall vor Augen zu haben, wenn er ein reines Morphemlexikon fordert.
31/2) Vgl. dazu besonders REIFLER (Compounds) mit dem allbekannten Beispiel KULTURINFILTRATION.
31/3) GUTER ist somit nur Positiv !

32/1) Vergleiche BIERWISCHS (Hierarchie) Teil-von-Relation und auch Möglichkeiten, wie sie NEUHAUS (Wortableitung) für die Ableitung von Adjektiven im Englischen aufgezeigt hat.

33/1) Vgl. dazu KLEIN (Analysegrammatik) S. 16 ff.
33/2) KLEIN (Studien), S. 126 ff. gelangt zu ähnlichen Ergebnissen.

35/1) Zu einer generellen Trennung in zwei Hauptgruppen kommt etwa TESNIERE (Esquisse) : in Leerwörter (mots vides) und Vollwörter (mots pleins),.
35/2) CHOMSKY (Aspekte) S. lo8.
35/3) Vgl. dazu HAYS (Computational Linguistics), bes. Kap. 5; ausfüljrlich dazu ebenfalls und den neueren.technischen Möglichkeiten Rechnung tragend WENZEL (Textverarbeitung), bes. Kapitel 3 S. 52 ff.; eine vergleichende Übersicht gibt auch - besonders orientiert an den Problemen des Information Retrieval – YANG (Search).

37/1) Hier sei zum besseren Verständnis eine technische Bemerkung angeführt: Der Erkennungsalgorithmus kann durchaus so aufgebaut sein, daß eine Textwortform nach allen Gesichtspunkten der Gliederung (Vollform, Stammform, Ableitung/Komposition 'parallel' untersucht wird, wenn dadurch Analysezeit - etwa mehrere Wörterbuchläufe nacheinander - eingespart werden kann. Die Hierarchieregel würde dabei erst angewendet, wenn im Gesamtergebnis mehrere Zuordnungen möglich erscheinen.
37/2) Dies gilt nicht für die Morphemebene; hier muß stets eine Priorität der Stamm- oder Wortformebene angenommen werden.

39/1) Zum Problem der Systematisierung der Wortnaht vgl. BOTHA (Lexicon), Kap. V und (allgemeiner) zum Verbindungsmorphem ZEPIC (Nominalkomposita).

39/2) Ich übernehme hier diesen Terminus von BÜNTING (Morphologische Strukturen), da er m.E. eine recht glückliche Abgrenzung gegenüber dem Begriff 'Stamm' darstellt: 'Kernmorpheme ... sind alle diejenigen Morpheme, die semantische Bedeutung haben und wortfähig sind ... w Unter S t a m m wird der Teil eines Wortes verstanden, der nach Abtrennung von Flexionsmorphemen übrig bleibt'. (S. 27 f.).

40/1) GRAFF (Sprachschatz), Vorrede zum ersten Band, S. I, der 1834 abgeschlossen wurde. Der 6. (letzte) Band erschien 1842; Graff hatte 1821 mit dem Werk begonnen.
40/2) GRIMM (Wörterbuch). Der letzte (16.) Band erschien 1954.
40/3) 3. Lieferung Februar 1968.

41/1) Vgl. zu diesem Problem ausführlich BAHR (Lexikographie). Hier genügt schon ein charakteristisches Zitat: 'Nach unsern Erfahrungen bietet erst eine Basis von einer Million Textseiten, verteilt auf etwa 5000 Quellen, die Gewähr dafür, daß aus ihnen der Wortschatz der deutschen Sprache sowohl dem Wortbestand als auch den Wortbedeutungen nach einigermaßen vollständig erfaßt werden kann.'(S. 98) Wenn dies auch auf ein historisches Wörterbuch bezogen ist, so ist diese Zahl doch beängstigend hoch.

41/1) Zum Problem des 'Veraltens' eines Wortes vgl. SCHOENE (Mots).
41/2) Man denke an Wortschatzssmmlungen wie das Advanced Learners Dictionary oder PFEFFER (Grunddeutsch).

43/1) CHOMSKY (Aspekte) S. 40 ff
43/2) Zum Komplex der Sprachentwicklung vgl. den Sammelband SMITH (Genesis).

44/1) Vgl. dazu KING (Grammar), der auf diesen historischen Aspekt im Hinblick auf die TG näher eingeht.
44/2) Ich will hier nur das Problem der Wortverdrehungen andeuten, wo-aus in der Regel bereits vorhandenen Morphemen neue Wörter ('MEINUNGSDUMMFRAGE', DOKTORKITTEL' ('TITEL')) gebildet werden, dabei die semantische Komponente ausnutzend.

45/1) Vgl. dazu auch Kap. 3.2.3.
45/2) Bei kleinerem Lexikonumfang - orientiert an sprachlichen Häufigkeiten - trifft dies nicht unbedingt zu, da die unregelmäßigen Bildungen gerade bei den häufigen Wörtern zu finden sind.
45/3) Vgl. z.B. die Zählungen von MEIER (Sprachstatistik), der (s. 53, Tabelle) für Wortformen ermittelt hat, daß die häufigsten Wortformen bei einem Lexikonanteil von 15,92 % (= 41.o83 verschiedene Wortformen) einen Textanteil von 96,29 % (= 10.505.939 Belege) erreichen.

49/1) SCHWEISTHAL (Präpositionen) S. 7, zählt 261 deutsche Präpositionen. Vgl. auch SCHMITZ (Präpositionen).
49/2)1Vgl. dazu eine Untersuchung von ROTHKEGEL (Syntagmen), die mir in Maschinenschrift vorlag.

51/1) Derartige Ausnahmen lassen sich verhältnismäßig leicht mithilfe rückläufiger Wörterbücher, etwa dem von MATER (Wörterbuch), erstellen.

52/1) Dies bezieht sich natürlich nur auf verschiedene Wörterbuchsuchläufe; bei einem Wörterbuchsuchlauf können die Regeln einmal auf alle glei•chlautenden Wortformen angewendet werden.

52/2) Zu dem Problem der richtigen Auswahl fachspezifischer Wörter vgl. GOUGENHEIM (Elaboration).

54/1) Nach einem derartigen Verfahren arbeitet etwa der Übersetzerdienst der westdeutschen Bundeswehr, der für die lexikalische Zuordnung einen Computer verwendet. Vgl. BERNER (Sprachmittlerwesen).

54/2) GRUBER (Relation) geht beispielsweise davon aus, daß den Lexikoneinträgen bestimmte prälexikale Merkmale beizugeben sind, die alle möglichen semantischen Relationen spezifizieren. Einige umfassen (ineorporate) komplexe Merkmale, so z.B. das Merkmal "FOOD" die Objekteigenschaften von EAT. Es wäre interessant, zu prüfen, inwieweit derartige semantische Relationsangaben bei der automatischen Klassifikation eines unbekannten Wortes mitverwendet werden könnten.

55/1) Zu den Problemen und Stufen des Lernens s. ZEMANEK (Automateh), bes. S. 1391 und 1396.

55/2) Eine liste mit Typisierung der Syntaktischen Mehrdeutigkeiten gibt WEBlsR (Homographie) anhand zahlreicher deutscher Beispiele. Eine allgemeine Typisierung versucht AGRICOLA (Polysyntaktizität).^

56/1) Zur Statistik (Erfolgsquote) vgl. WEBER (Homographie), S. 85 ff.

58/1) Zum Problem der Eigennamen vgl. noch unten Kap. 3.2.4

58/2) Vgl. die (maschinellen) Studien von GEENS, der die gesamte Graphemstruktur eines Wortes zur automatischen Wortklassenreduktion und -erkennung heranzieht (mündliche Mitteilung). Zu dem Analyseprojekt in Leuven (Belgien), an dem GEENS mitarbeitet, vgl. ENGELS (Analyse).

59/1) Vgl. dazu das ausführliche Reduktionssystem in Kap. 4.2.2 2 Näheres s. die Statistik S.103 und die Tabellen lV und V des Anhangs.

60/1) Die Wahrscheinlichkeit einer Lösung ist etwa abhängig von der unterschiedlichen token-Häufigkeit, also von der rela-. tiven Häufigkeit der einzelnen Lösungsmöglichkeiten.

62/1) Diese Auffassung untersdheidet sich von der LENDERS: (Lexical Systems), der ein dynamisches (Computer-) Lexikon der Art definiert hat, daß alle Bestandteile wandlungsfähig sein sollen. Als Beispiel dient ihm das geplante Kumulative Wörterbuch des Instituts für Kommunikationsforschung und Phonetik in Bonn. (Vgl. dazu KRALLMANN (Datenbank)). In wesentlichen weiteren Punkten ist jedoch der Auffassung Lenders von einem wandlungsfähigen, text- und themabezogenen maschinellen Lexikon zuzustimmen.

63/1) Vgl. zu dem Problem der automatischen Erkennung von Eigennamen noch BORKOWSKI (Personal Names).

64/1) Damit lassen sich die oben angeführten Mehrdeutigkeiten allerdings noch nicht lösen.

67/1) Vgl. unten S. lo3,

67/2) Für eine ausführliche Beschreibung der Syntaxanalyse verweise ich auf den entsprechenden Bericht: EGGERS y(Syntaxanalyse). Das hier zugrundegelegte Verfahren ist bereits seit l970 abgeschlossen; das mittlerweile in Saarbrücken begonnene Projekt 'Automatische Lemmatisierung' wird ebenfalls eine Analyse zugrundelegen; diese Analyse II liegt jedoch noch nicht in Programmform vor.

67/3) Vgl. hierzu besonders EGGERS (Sprache).

69/1) EGGERS (Syntaxanalyse).

69/2) Vgl. EGGERS (Syntaxanalyse) S. 40 ff.

69/3) Etwa in der Richtung, wie sie für das Projekt 'Automatische Lemmatisierungl gegenwärtig erstellt worden ist (Siehe LEMMATISIERUNG).

70/1) Der zur Analyse damals zur Verfügung stehende Computer, eine Electrologica ä1, wäre zudem einem technisch so aufwendigen Dictionary-Look-up, wie er zur Ermittlung präfigierter Wörter oder zur Komposita-Erkennung notwendig ist, nicht gewachsen gewesen, da als periphere Speicher nur Magnetbänder zur Verfügung stehen, deren Spulzeiten in diesen Fällen unakzeptabel hoch gewesen wären.

70/2) Näheres siehe wieder: EGGERS (Syntaxanalyse) S. 55 ff. Vgl. dazu auch HEINRICH (Wörterbuch).

72/1) Vgl. EGGERS (Syntaxanalyse) S. 62 ff. und ZIMMERMANN (Mehrdeutigkeiten) S. 39 ff.

73/1) Das Analysesystem gewährleistet natürlich, daß die unabhängig von einer spezifischen HO-Lösung lösbaren Homographen zuerst gelöst werden.

73/2) EGGERS (Syntaxanalyse) S. 85 ff.

75/1) Die Satznummern beziehen sich auf die Testsätze des Anhangs; aus ökonomischen Gründen wurde die.2 der Zehntausenderstelle weggelassen, da sie nur zur hier nicht interessierenden Unterscheidung der rde- und FAZ-Sätze von diesen Testsätzen dient.

75/2) Vgl. dazu die Tabelle I des Anhangs S.'121.Zu den durch die Suffixanalyse und das Zahlenerkennungsprogramm klassifizierten Wörtern siehe auch Tabelle III des Anhangs S.431. Zur Auswertung der Testsätze im Hinblick auf die morphologische Klassifizierung vgl. S. Ao4 {

75/3) Es wird natürlich vorausgesetzt, daß Wörter wie KURZ, HERZ, SCHUTZ, SCHATZ als'Ausnahmerf im Lexikon stehen.

78/1) Zu Beispielen dafür aus den Testsätzen vgl. unten Kap.4.3.3.2.

80/1) Vgl. aber die Restriktionen für mögliche generelle Mehrdeutigkeiten S.449F.

80/2) Die Wortklasse NAM ist zwar in das Lexikon und das entsprechende Verfahren integriert, doch ist noch kein entsprechendes Homographenauflösungsprogramm vorhanden, so daß keine Lösung NAM auftritt.

80/3) Diese Zusätzangaben sind jedoch nicht Gegenstand desl'Ests, da die entsprechenden Routinen technisch noch nicht völlig ausgearbeitet bzw. getestet sind.

81/1) Beim Adjektiv ist in den Bits 22 - 24 der Endungs/Deklinationstyp angegeben: 1 = E, 2 = EN, 3 = EM, 4 = ER, 5 = ES.
81/2) In den Bits 22 - 24 der Verbzelle stehen die morphologisch ermittelten Konjugationsangaben: 1 = E, 2 = ST, 3 = T/ST 4 = N.
81/3) In den Bits 22 - 24 der Infinitivzelle ist der Zähler für IZU (Infinitiv mit 'ZU' im Wort) untergebracht.
81/4) Die Zähler der Rektionsangaben haben als möglichen Maximalwert die Zahl 15

82/5) Die Zähler der Wortklassenzellen und der SEIN/HABEN-ZELLE können theoretisch den maximalen Wert 31 erreichen.
82/6) Zum Begriff der Analyseeinheiten Vgl. EGGERS (Syntaxanalyse) S. 9o ff. Hier nur ein verdeutlichendes Beispiel mit drei Analyseeinheiten: DER MANN / DER DORT GEHT / IST MEIN FREUND.
82/7) Homographendurchlaufzähler '- 6
82/8) Beispielsweise ist die Lösung des Wortes KOSAKEN in Satz 9 als 'unsicher' gewertet worden, da es erst im 9. Durchlauf in Abhängigkeit von der noch nicht gelösten mehrdeiçtigen Wortform WAREN (VRB/INF/SUB) klassifiziert wurde.<Anm.: Die Fußnoten zu S. 81 und 82 befinden sich alle auf S. 82.>

83/1) Weitere empirische Tests könnten das Ergebnis bei veränderten Werten durchaus verbessern. Vgl. Kap. 4.3.3.4.
83/2) Näheres wird dazu in Kap. 4.2.3 ausgeführt.

84/1) Näheres dazu wieder bei der Beschreibung der Testergebnisse.
84/2) Beispiele zu den Typen - die alle auch bei natürlichen Homographen belegt sind - finden sich in EGGERS (Syntaxanalyse) S. 67 ff; Beispiele für unbekannte Wörter geben die Tabellen IV und V des Anhangs.

85/1) = der für die quasimorphologische Analyse verwendete Typ.
85/2) Homographentyp der Syntaxanalyse.
85/3) Konjugationstyp (vgl. Anm.2 5.92 .
85/4) Deklinaüonstyp (vgl. Anm.1 5.92-).
85/5) Bei SUB ist zugleich NAM eingeschlossen.
85/6) Diese Mehrdeutigkeit wird vorläufig in der Analyse automatisch zu SUB umgedeutet, da ein Homographenauflösungsprogramm für den Typ SUB/NAM noch nicht vorliegt.
85/7) Hierunter fallen alle durch die morphologische Analyse ermittelten Suffixe, da die Steigerung auf-ER prinzipiell vorzusehen ist, aber sprachlich nicht notwendig realisiert sein muß.
85/8) Hierzu gehören die Endgrapheme B, C, D, F, G, H, K, L, M (außer EM), P, R, S, V während die letzten Buchstaben J, Q, W, X, Y und Z unmittelbar zu NAM führen.
85/9) Hier ist nur der substantivierte Infinitiv generell möglich.

86/1) Faktisch war das nicht gegeben, daher resultieren auch einige Fehler des Tests.

87/1) Vgl. dazu S. 58 Anm. 2.
87/2) Vgl.. STOLZ (Coding) und KLEIN (Coding).
87/3) Näheres siehe STOLZ (Coding) und WISSYN.

88/1) Der Vollständigkeit halber sei bemerkt, daß dem Verfasser diese Modelle erst bekannt wurden, als er sein Verfahren schon zu entwickeln begonnen hatte.
88/2) Beispielsweise brachte das WISSYN-Verfahren eine Quote von über 9o % richtiger Lösungen. Vgl. STOLZ (Coding) S. 404 f.

89/1) So wird etwa zum Leidwesen vieler Schüler der 'Subjonctif' des Französischen sehr ausführlich behandelt, obgleich er aus der französischen Umgangssprache fast völlig verschwunden ist.

90/1) Die in dem vom System her nicht mehr entscheidbaren Fällen herangezogene statistische Wahrscheinlichkeit ist also nicht prinzipiell systemorientiert, sondern eine vorläufige Kapitulation vor der Komplexität der Regel. Vgl. EGGERS (Syntaxanalyse) S. 80.

91/1) Vgl. die Liste der Typen in EGGERS (Syntaxanalyse) S. 65 ff.

92/1) Vgl. den Probeabdruck S.µ9.
92/2) Punkt, Fragezeichen und Ausrufezeichen dienen als Erkennungszeichen für das Satzende.

93/1) WB = 'normales' Wortformenbuch; KL = Kurzzeitlexikon.

97/1) Dies hatte auch einen technischen Grund: Für alle Testpersonen wurde jeweils ein Analysezyklus durchgeführt. So war die Gewähr gegeben, daß bei jeder Wörterbuchsuche die Ergebnisse des vorherigen Zyklus bei den unbekannten Wörtern im Kurzzeitlexikon abgefragt werden konnten.
97/2) An dieser Stelle möchte ich den elf Testpersonen für ihre Mühe herzlich danken, zunächst meiner Frau Ellen und meinem Freund Gerd Szielasko, die sich dabei einer für sie ungewohnten Aufgabe unterziehen mußten; nicht weniger dankbar bin ich den Damen A. Rothkegel, L. Schmidt und R. Weber sowie den Herren R. Dietrich, W. Klein, H.D. Maas, J. Neuhaus, R. Rath und H.J. Weber für ihre freundliche Unterstützung.

98/1) Zur Aufschlüsselung der Wörter nach morphologischen Gegebenheiten und zur Homographentypisierung siehe Tab. II.
98/2) Aufgrund eines zu spät entdeckten Ablochfehlers (Punkt und Ausrufezeichen unmittelbar hintereinaüer) läuft die Satz zählung bis 283; die Satznummer 156 entfällt.
98/3) Die Reihenfolge der Tabelle ist nicht identisch mit der Namenfolge in Anm. 2 S. 97.

99/1) Satznummer 156 entfällt.
99/2) Vgl. Tabelle XIV des Anhangs.
99/3) Vgl. darüberhinaus die Angaben zur Satzlänge der Zeitungstexte Kap. 4.3.3.3.

100/1) Vgl. die Analyse von Satz 4o im Anhang.

101/1) Satzanalyse, im Anhang
101/2) Vgl. dazu Tabelle III des Anhangs.

102/1) Die Zahlen in Klammern bezeichnen die Satznummern.

103/1) Satz 265: WIR SAHEN ETWAS DUNKELBRAUNES, UNDEFINIERBARES SICH LANGSAM AUF UNS ZU BEWEGEN.
Satz 23: ALS DIE DREI IM FAHRSTUHL SICH WUNDERTEN, DASS NACH DEM ZEHNTEN GESCHOSS NICHT, WIE ERWARTET, DAS ELFTE FOLGTE, GING ES SCHON BERGAB. (ELFTE ist als Substantiv gelöst).

105/1) Vgl. dazu ausführlich EGGERS (Syntaxanalyse) S. 83 ff.
105/2) Es handelt sich um die nicht im Lexikon - wie postuliert - aufgenommenen Partizipien Starker Verben: BESONNEN (Satz 6) und AUFGERUFEN (Satz 235).

106/1) Vgl. EGGERS (Syntaxanalyse) S. 87 f.
106/2) Es ist übrigens interessant, daß eine gegebene Wortform nicht ein einziges Mal fälschlich als Infinitiv bestimmt wurde, obgleich diese Möglichkeit immerhin in 69 Fällen (HO 14-en und HO 14-n) theoretisch gegeben war.
106/3) Vgl. EGGERS (Syntaxanalyse) S. 87 f.

107/1) Anzahl der Deutigkeiten in römischen Ziffern.
107/2) Hier nicht mitgerechnet sind die 7 Belege von H0 55 und die drei Belege zu HO 4, 8 und 20.
107/3) Bei der Auswertung des Tests sind die Lösungen der entsprechenden natürlichen Homographen nicht berücksichtigt worden, da sie in diesem Zusammanhang ohne Bedeutung sind.

108/1) Vgl. EGGERS (Syntaxanalyse) S. 88 f.

109/1) Vgl. EGGERS (Syntaxanalyse) 3. 88 f
109/2) Analyse des Satzes im Anhang.
109/3) Weitere Beispiele: Satz 9o, 142, 143, 149, 153, 177...
109/4) Analyse des Satzes im Anhang,

110/1) Weitere Bedspiele in den Sätzen 138,183,195,2o2,253,...
110/2) Eine 'feinere' Analyse, die die M3glichkeit böte, die Ellipse zu erkennen, könnte hier die Fehlerquote entscheidend veaingern.

111/1) Weitere Beispiele in den Sätzen 149, 172, 191, 194,...
111/2) Ein derartiges System ist für die möglichst fehlerlose Ergebnisse verlangende Automatische Lemmatisierung heute wohl noch unumgänglich; es ist daher auch geplant, es in die Analyse II des Saarbrücker Projekts zu integrieren.Vgl. LEMMATISIERUNG S. 14.

112/1) Saarbrücker Zeitung (SZ) vom 27.7.1971.
112/2) FÜR SIE Nr. 16 (1971).
112/3) DER SPIEGEL Nr. 28 (1971).
112/4) Zur Unterscheidung von Testsätzen mit vorgegebenen Wörtern beginnen diese Satznummern bei 500.

113/1) Nur die ersten Sätze eines jeden Artikels wurden herangezogen.
113/2) Einbezogen ist hier die Anwendung der Substantivierungsregel.

113/3) Vgl. S.102.
113/4) Vgl. zum Einzelnen die Tabelle XV des Anhangs. Zum Vergleich die Quoten der Analyse der Testsätze mit vorgegebenen Wortformen (Test 1) und des Tests der Syntaxanalyse.

115/1) Die falsch gelösten Substantive machen etwa 6o % aller Analysefehler aus. Hier zeigt sich vielleicht der Wert der Großschreibung des Anfangsbuchstabens.bei Substantiven; eine Verwertung dieser Information hätte die Analyseergebnisse drastisch verbessert: Viermal wurde zudem fälschlich SUB für eine andere Wortklasse ermittelt.
115/2) Im Zusammenhang mit der morphologischen Analyse ist das häufige Auftreten von Zahlen erwähnenswert, die den Wert des Zahlenerkennungsprogramms durchaus reohtfertigen.

116/1) Die Ausnahmen AUFGEHOBEN und VORAUSGESEHEN wurden oben behandelt.
116/2) Einzelheiten siehe wieder in Tabelle XV. Beispielsätze für die Analyse (503, 530, 539, 547) im Anhang.

117/1) Durch den Druckfehler BUNBESPOLITIKER, der zu spät bemerkt wurde, ist das Wort in einem Zyklus nicht vertreten.
117/2) Vgl. Tabelle IV.

118/1) Es ist natürlich denkbar, daß man hier die Lösungen VRB mit der Lösung SUB kombiniert, da bei VRB und Endung EN stets eine (generelle) Substantivierungsmöglichkeit mitgegeben ist.
118/2) Ich klammere im folgenden bei allen Reduktionen das 'Verbieten von NAM aus, da dafür aus schon erwähnten Gründen kein Beleg zu erwarten war.
118/3) Bei mehreren noch möglichen Klassen ist die Wortklasse, die die Endphase erreicht hat, unterstrichen.
118/4) Hier ist ADV zwar formal verboten, aber generell noch zugelassen.

120/1) In der Analyse einschließlich Kurzzeitlexikon ist ein derartiger Fall belegt. So geschah es, daß EINGELOEST – obgleich ADV bereits verboten war – in Satz 256 als HO 9 (leider fälschlich) als ADV gelöst wurde, während es als HO 10 bei der Analyse ohne Kurzzeitlexikon richtig als PTZ bestimmt worden war.
121/1) Insgesamt war 'unsicher' nur in 28 Fällen aufgetreten. Bei 11 Belegen war die Lösung in der Tat falsch. Jeweils einmal waren die Lösungen für die folgenden *vorgegebenen* Wortformen 'unsicher': ANLAECHELN AXT BEDECKEM (f) BEGAFFT DUNKELGRUENEN FELSENFEST (f) GRINST KOSAKEN.

122/1) Die Wortklasse NAH ist wieder weggelassen. In Klammern stehen die möglichen richtigen Lösungen. Das Zeichen * zeigt an, daß diese Wortklasse aufgrund der quasimorphologischen Analyse von vornherein ausgeschlossen wurde.
122/2) Hier ist jeweils eine 'unsichere' Lösung verzeichnet.
122/3) In Satz 220 ist DUNKELBRAUN aufgrund eines Programmierungsfehlers fälschlich als ADJ gelöst, obwohl diese Wortklasse bereits im Lexikon ausgeschlossen war.

126/1) Vgl. QUILLAN (Semantic Memory).

127/1) Vgl. TOLLENAERE (Lexikographie).
127/2) Human ist hier im Gegensatz zu ‚maschinell' zu verstehen.
1273) Vgl. LENDERS (Lexical Systems).

129/1) Maschinenintern sind die hier unter den unflektierten Formen verzeichneten Suffixe nach ihren möglichen Flexionsformen aufgelöst, also z.B. IG zu IGE, IG, IGES, IGEM, IGEN, IGER. Daneben sind die dann noch möglichen Deklinationsangaben vermerkt.
129/2) Hier ist nur die Endung EM zugelassen, um nicht bei Verben wie ERREICHEN oder Substantiven wie HIMMELREICH zu Fehlern zu gelangen.
129/3) Unflektiert: ADV; flektiert:ADJ.

130/1) Die Wortformen wurden so ausgewählt, daß bei jedem Homographen jede Deutigkeit mindestens durch ein richtiges Resultat vertreten war.
130/2) Hierzu gehören (außer den oben und den in der Liste zur morphologischen Reduktion verzeichneten Fällen) die Endgrapheme B,C,D,F,G,H,K,L,M,P,R,S,V.

145/1) Ergebnis-Wortklasse nach der Analyse
145/2) Kriterium (Graphemfolge) für die morphologische Präklassifikation
145/3) Homographen-Typ
145/4) Bei falscher Klassifizierung: richtige Wortklasse
145/5) 'Wortformen', die vor dem ersten Lexikoneintrag stehen, werden als (Zahl)Adjektive betrachtet. Die Substantivlösungen ergeben sich nach Anwendung der Substantivierungsregel.

150/1) Vgl. dazu im Einzelnen EGGERS (Syntaxanalyse).
150/2) Erklärung der Abkürzungen dieser Rubrik S. 71.
150/3) NOG = Nominalgruppe

151/1) Vgl. EGGERS (Syntaxanalyse).

<S. 128>

ANHANG

<S. 129>

TABELLE I: Suffixliste zu den morphologischen Redundanzregeln[129/1]

Suffix (Endgraphem)	*Wortklasse*	*Suffix (Endgraphem)*	*Wortklasse*
AL	ADJ/ADV[129/3]	MAL	ADV
AND	SUB	MENT	SUB
BAR	ADJ/ADV	NIS	SUB
EI	SUB	OLOGE	SUB
EIL	ADJ/ADV	OS	NAM
END	ADJ/ADV	Q	NAM
ENZ	SUB	REICHEM[129/2]	ADJ
HAFT	ADJ/ADV	SAM	ADJ/ADV
KEIT	SUB	TAFT	SUB
IA	SUB	TION	SUB
IAS	NAM	TIV	ADJ/ADV
IE	SUB	TUM	SUB
IG	ADJ/ADV	UNG	SUB
1K	SUB	UR	SUB
ISCH	ADJ/ADV	US	SUB
ISMUS/EN	SUB	VOLL	ADJ/ADV
IST	SUB	W	NAM
IUM	NAM	WEISE	ADJ/ADV
J	NAM	WERT	ADJ/ADV
KEIT	SUB	X	NAM
LEI	ADJ	Y	NAM
LICH	ADJ/ADV	Z	NAM
LOS	ADJ/ADV		

<S. 130>

TABELLE II: Vorgegebene unbekannte Wortformen[130/1]

Quasimorph. Analyse	Homographen-Typ	richtige Lösung	Wortform
ES	5	ADJ	DUNKELBRAUNES
ES	5	SUB	MACHTKOMPLEXES
EM	5	ADJ	DUNKELBRAUNEM
EM	5	SUB	PHONEM
ST	10	VRB	GRINST
ST	10	PTZ	EINGELOEST
ST	10	SUB	PEST
ST	10	ADV	FELSENFEST
T	10	VRB/PTZ	BEGAFFT
T	10	SUB	AXT
T	10	VRB	KRAXELT
E	14	ADJ	DUNKELBRAUNE
E	14	VRB	VERBEUGE

E	14	SUB	ANGSTHAST
EN	14	ADJ	DUNKELGRUENEN
EN	14	ADJ/VRB	BEANSTANDETEN
EN	14	SUB	KOSAKEN
EN	14	VRB/INF	BEDECKEN
N	14	ADV	DUNKELBRAUN
N	14	VRB/INF	ANLAECHELN
ER	46	ADV/ADJ	GELBER
ER	46	SUB	BUNDESPOLITIKER
ER	46	ADJ	ANALOGER
Rest[130/2]	11	ADV	BERGAB
"	11	SUB	ANLAUFS
	11	SUB	ABDRUCK

<S. 131>

Tabelle III Morphologisch klassifizierte Wortformen

Wortlaut	*richtig*	*falsch*	*Satznummer*	*Suffix*
ABSTOSSENDES	ADJ		163	ENDES
ALLWISSENDEN	ADJ		131	ENDEN
ANALYSIERBAR	ADV		85	BAR
AUFFAELLIG	ADV		201	IG
BILDUNGSPOLITIK	SUB		247	IK
BUNDESPOLITIK	SUB		263	IK
CIRCUS	SUB		269	US
DEIMOS	NAM		116	OS
DEMONSTRIERENDEN	ADJ		15	ENDEN
DIFFERENTIALGLEICHUNGEN		SUB	213	UNGEN
ELEGANZ	(SUB)	NAM	239	Z
ELFTE	(ADJ)	SUB	23	(ZAHL)
ERSCHWINGLICHEN	ADJ		7	LICHEN
FRAUENBEKANNTSCHAFTEN	SUB	(ADJ)	163	HAFTEN
FUENF	ADJ		26	(ZAHL)
GEMAELDEAUSSTELLUNG	SUB		216	UNG
GEROELL	SUB		136	ELL
GESCHAEFTSGEBAREN	SUB	(ADJ)	197	BAREN
GESCHMACKLOSIGKEIT	SUB		283	KEIT
GRUNDEINHEITEN	SUB		179	HEITEN
HALTERUNGEN	SUB		134	UNGEN
HINDERNISSE	SUB		215	NISSE
HINTERLASSENDEN	ADJ		181	ENDEN
HINZUFUEGUNG	SUB		207	UNG
HOFFNUNGSVOLLE	ADJ		214	VOLLE
HYGIENISCHEN	ADJ		198	ISCHEN
LANGJAEHRIGEN	ADJ		222	IGEN
NEUNZEHNHUNDERTSIEBZIG		ADJ	220	(ZAHL)
OHRFEIGE	SUB	(ADJ)	14	IGE
OPPOSITIONSPARTEIEN	SUB		217	EIEN
PETRUS	SUB		68	US
PFLICHTUEBUNGEN	SUB		34	UNGEN
PHOBOS	NAM		116	OS

"	"	117	“
PHONOLOGIE	SUB	229	IE

<S. 132>

Wortlaut	*richtig*	*falsch*	*Satznummer*	*Suffix*
PHONOLOGISCHE	ADJ		137	ISCHE
POLIZEISTELLE	SUB	(ADJ)	101	ELLE
RASSISMUS	SUB		4	MUS
REBROW	NAM		218	W
SCHAL	SUB		250	AL
SCHMIEDEND	ADV		174	END
SINNUNTERSCHEIDENDE	ADJ		21	ENDE
SKALEN	SUB		137	ALEN
SPIEGELSAAL	SUB		12	AL
STRASSENKREUZUNG	SUB		141	UNG
STRUKTURALISTEN	SUB		178	ISTEN
TIEFERLIEGENDEN	ADJ		140	ENDEN
TOENUNG	SUB		115	UNG
UNDEFINIERBARES	SUB	(ADJ)	265	BARES
UNENTGELTLICH	ADV		207	LICH
UNHEILVERHEISSEND	ADV		219	END
UNORDENTLICHEN	ADJ		179	LICHEN
UNZERSTOERBAR	ADV		224	BAR
VERHEERENDEN	ADJ		15	ENDEN
”	”		222	”
VERLEIHUNG	SUB		209	UNG
VERSCHNUERUNG	SUB		134	UNG
VERSPRECHUNGEN	SUB		274	UNGEN
VERZEICHNIS	SUB		101	NIS
WUCHTIGEM	ADJ		174	IGEM
ZISCHEN	SUB	(ADJ)	218	ISCHEN
ZWEIFELHAFTEM	ADJ		96	(ZAHL)

<S. 133>

TABELLE IV Syntaktisch klassifizierte unbekannte Wortformen

I: Vorgegebene Wortformen

HO	*Wortlaut*	*r*	*f*	*richtig*	*falsch*	*a*	*Satznummern*
11	ABDRUCK	11	-	SUB	-	11	24,50,76,101,128,133, 181,207,231,259,282
46	ANALOGER	10	1	ADJ	-	10	30,56.81,108,145,161, 187,213,249,264
				(ADJ)	SUB	1	4
14-e	ANGSTHASE	8	4	SUB	-	8	26,78,102,103, 183,209,241,282
				(SUB)	ADV	2	52,130
				(SUB)	VRB	2	138,234
14-n	ANLAECHELN	9	3	VRB	-	5	12,38,116,221,245

				SUB	.-	4	64,92,272,272
				(SUB)	ADV	1	195
				(INF)	SUB	1	149
				(INF)	VRB	1	169
11	ANLAUFS	10	1	SUB	-	10	22,48,74,99,126,144, 179,205,230,282
				(SUB)	ADV	1	236
10	AXT	9	2	SUB	-	9	17,43,69,95,121,174, 225,238,277
				(SUB)	ADV	1	143
				(SUB)	PTZ	1	200
14-en	BEANSTANDETEN 8,34,60,88,135		10	2	ADJ	-	5
				VRB	-	5	112,134,217,251,268
				(ADJ)	VRB	1	165
				(VRB)	SUB	1	191
14-en	BEDECKEN	7	4	INF	..	1	10
				VRB	-	6	36,62,114,167,193,252
				(INF)	SUB	1	270
				(VRB)	SUB	2	go,219
				(VRB)	ADJ	1	142
10	BEGAFFT 18,44,70,122,226,258,276	9	2	VRB	-	7	
				PTZ	-	2	84,201
				(VRB)	SUB	1	141
				(VRB)	ADV	1	175

<S. 134>

HO	*Wortlaut*	*r*	*f*	*richtig*	*falsch*	*a*	*Satznummern*
11	BERGAB	10	2	ADV	-	10	23,49,75,100,127,136, 180,206,232,236
				(ADV)	SUB	2	260,282
46	BUNDESPOLITIKER		8	2	SUB	-	8
							3,29,55,146,160,186, 247,263
				(SUB)	ADJ	1	81
				(SUB)	ADV	1	107
14-n	DUNKELBRAUN	5	6	ADV	-	3	11,63,91
				SUB	-	2	168,271
				(ADV)	VRB	3	37,153,194
				(ADV)	SUB	1	115
				(SUB)	ADJ	1	220
				(SUB)	ADV	1	253
14-e	DUNKELBRAUNE	11	-	ADJ	-	9	25,51,77,102,129,208, 233,261,283
				SUB	-	2	151,182
5-em	DUNKELBRAUNEM		10	1	ADJ	-	9
							20,46,72,124,154,203, 228,239,280

				SUB	-	1	177
				(ADJ)	SUB	1	97
5-es	DUNKELBRAUNES	9	2	ADJ	-	7	5,31,57,83,188,214,249
				SUB	-	2	162,265
				(ADJ)	SUB	2	109,157
14-en	DUNKELGRUENEN		11	1	ADJ	-	11 7,33,59,86,86,111, 158,190,216,250,267
				(ADJ)	SUB	1	164
10	EINGELOEST	10	1	PTZ	-	10	14,40,66,93,118,135, 197,223,256,274
				(PTZ)	SUB	1	171
10	FELSENFEST	3	8	ADV	-	3	68,139,255
				(ADV)	VRB	3	16,94,120
				(ADV)	SUB	4	42(meta),173,199,224
				(SUB)	ADV	1	276
46	GELBER	9	3	ADJ	-	7	1,28,54,185,211,246, 262
				ADV	-	2	2,106

<S. 135>

HO	*Wortlaut*	*r*	*f*	*richtig*	*falsch*	*a*	*Satznummern*
				(ADJ)	SUB	1	80
				(ADV)	SUB	2	152,159
10	GRINST	11	-	VRB	-	11	13,39,65,87,117,155, 170,196,222,242,273
14-en	KOSAKEN	4	7	SUB	-	4	9,89,113,218
				(SUB)	VRB	5	35,61,166,244,269
				(SUB)	ADJ	1	132
				(SUB)	ADV	1	192
10	KRAXELT	11	1	VRB	-	11	19,45,71,96,123,148, 176,176,202,257,279
				(VRB)	PTZ	1	227
5-es	MACHTKOMPLEXES		10	1	SUB	-	10
							6,58,85,110,140,163, 189,215,243,266
				(SUB)	ADJ	1	32
10	PEST	10	1	SUB	-	10	15,41,67,105,119,147, 198,222,254,275
				(SUB)	VRB	1	172
5-em	PHONEM	12	-	SUB	-	12	21,47,73,98,125,137, 137,178,204,229,237,2 281
14-e	VERBEUGE	11	-	VRB	-	11	27,53,79,104,131,150, 184,210,235,240,283

<S. 136>

TABELLE V Syntaktisch klassifizierte unbekannte Wortformen II: Nicht vorgegebene Wortformen

HO	*Wortform*	*richtig*	*falsch*	*Satznummer*

14-e	AEXTE	SUB		95
14-e	AFFE	SUB		226
14-e	AFFENHERDE	SUB		227
14-e	ALLERBESTE	SUB		159
11	ALLSEITS	ADV		7
14-en	ALKOHOLBEDINGTEN	ADJ		161
14-e	ANFANGE	(VRB)	SUB	83
14-en	ANHEIZEN	SUB		205
14-en	ANZUSTACHELN	(INF)	VRB	103
11	ASCH (metaspr.) (?)	ADV		165
14-e	ASCHE	SUB		10
14-en	AUFGERUFEN (Lex.)	(PTZ)	VRB	235
10	AUSGERUHT	PTZ		19
14-en	AUSSCHAUEN	(INF)	VRB	116
11	BEDEUTUNGSUNTERSCHIED	SUB		204
10	BEIGEMENGT	PTZ		280
46	BELIEBTER	ADJ		243
10	BENIMMT	VRB		238
11	BERGAUF	(ADV)	SUB	260
14-e	BERGSPITZE	SUB		96
5-es	BERUEHMTES	ADJ		216
14-en	BESONNEN (Lex.)	(PTZ)	ADV	6
14-n	BESUCHERIN	SUB		9
46	BIER	SUB		109
4-e	BLOESSE	SUB		167
5-em	BLONDEM	(ADJ)	SUB	177
46	BLUMENFELDER	SUB		90
5-em	BRUENETTEM	(ADJ)	SUB	177
55	BUDA	SUB		275
46	BUNDESPOLITIKER	SUB	ADJ	212
14-e	CAFE	(SUB)	VRB	169
"	"	SUB		258
10	CHARMANT	(ADV)	VRB	149
46	COMPUTER	SUB		109
„	"	SUB		131
14-en	DIENSTJAHREN	SUB		26
5-es	DUNKELGRUENES	ADJ		157
11	EHRGEFUEHL	SUB		184
10	EINGEDECKT	PTZ		86
10	ENTFUEHRT	PTZ		214
14-n	ENTVOELKERN	INF	SUB	147
55	EO (IPSO)	(ADV)	SUB	78
14-en	ERBITTERTEN	ADJ		166
14-n	ERWIN	SUB		172
11	ESEL	SUB		168
11	EULENBAUM	SUB		52
14-e	FAEHRTE	SUB		129
11	FAHRSTUHL	SUB		23

HO	*Wortform*	*richtig*	*falsch*	*Satznummer*

10	FEDERBETT	SUB		193
14-en	FEINDEN	SUB		166
14-n	FERNSEHZUSCHAUERN	(SUB)	ADV	202
14-e	FINSTRE	ADJ		174
14-en	FLECHTEN	SUB		114
	"	SUB		119
14-e	FLINTE	SUB		43
11	FORMULAR	SUB		264
14-e	FRISEURE	SUB		10
5-es	FUSSES	SUB		259
10	GARTENFEST	SUB	VRB	276
11	GELB	(ADV)	SUB	46
14-en	GEMUSTERTEN	ADJ		250
5-em	GENERATIONENPROBLEM	SUB		4
14-e	GEPEITSCHTE	(ADJ)	VRB	220
10	GERAUCHT	PTZ		5
14-e	GERUEHMTE	ADJ		7
10	GESCHAEMT	PTZ		6
11	GESCHOSS	SUB		23
11	GEWAECHS	SUB		121
14-en	GLOCKEN	SUB		2
14-n	GUTSCHEIN	SUB		93
14-e	GUTSCHEINE	SUB		256
11	HANDSCHUH	SUB		181
46	HEERFUEHRER	SUB		6
10	ELFT	VRB		13
14-en	HELLBLONDEN	ADJ		208
14-e	HELLERE	ADJ		115
14-en	HELLGRUENEN	ADJ	SUB	164
10	HERVORRUFT	(VRB)	SUB	204
14-e	HERZLICHE	ADJ		11
5-es	HOCHGEBIRGES	SUB		90
11	HOHL	ADV		59
46	HOLZFAELLER	(SUB)	ADV	174
11	HOTELS	SUB		227
14-en	HUENDCHEN	(SUB)	ADJ	57
10	HUETTENSTADT	SUB		246
4	IMITIEREN	INF		218
14-en	IMPFEN	INF)	VRB	254
14-en	JAHRMILLIONEN	SUB		106
14-en	JUGENDTAGEN	(SUB)	VRB	66
46	KAFFEEPULVER	SUB		5
14-n	KIESELN	SUB		279
11	KIRCHTURM	SUB		19
11	KNEF	SUB		180
14-en	KNOSPEN	SUB		252
46	KOHLENHAENDLER	(SUB)	ADV	10
5-em	KOSTUEM	SUB		218
”	”	SUB		250
46	KRANKHEITSERREGER	SUB		118
55	KRIPO	SUB		259
14-en	KUFEN	SUB		205

<S. 138>

HO	*Wortform*	*richtig*	*falsch*	*Satznummer*
14-e	LANDESSITTE	SUB		150
11	LEBERSCHWUND	SUB		161
14-e	LIEBLINGSFARBE	SUB		253
14-en	MASSGESCHNEIDERTEN	ADJ		91
11	MINIROCK	SUB		258
46	MISTER	(SUB)	ADJ	52
10	MITGESPIELT	PTZ		94
14-en	MITTELGROSSEN	ADJ		9
14-e	MODEFARBE	(SUB)	ADV	220
46	MODESCHOEPFER	SUB		203
14-en	MONATELANGEN	ADJ		88
14-e	MONDE	SUB		116
5-em	MORPHEM	SUB		178
14-en	NARZISSEN	SUB		252
14-n	NEGERN	SUB		11
14-en	NELKEN	SUB		31
10	OSTERZEIT	SUB		2
14-e	PERUECKE	SUB		13
14-e	PFOTE	SUB		128
5-em	PHONEMSYSTEM	SUB		98
14-e	PLANE	(SUB)	ADJ	270
14-e	PORSCHE	SUB		262
14-e	PROGNOSE	(SUB)	ADJ	203
14-en	PROTESTIERTEN	VRB		103
11	PUFF	SUB		169
14-en	PUPILLEN	(SUB)	ADJ	153
14-e	PURZELTE	VRB		144
14-en	REITEN	(VRB)	SUB	269
14-n	RENNRODELN	SUB		205
14-en	REPARATURARBEITEN	SUB		88
10	RUNTERFALLT	VRB		71
14-e	SAEGE	(SUB)	ADV	143
55	SAFARI	(SUB)	ADV	254
10	SAFT	(SUB)	PTZ	97
11	SCHECK	SUB		171
”	“	SUB		223
14-n	SCHIMMERN	(VRB)	SUB	153
5-es	SCHLICHTES	ADJ		188
11	SCHMETTERLING	SUB		28
46	SCHNEEFELDER	(SUB	ADJ	90
14-en	SCHNEEFLOCKEN	(SUB)	VRB	142
14-en	SCHNEESCHICKTEN	SUB		252
14-e	SCHUHCREME	SUB		261
14-en	SCHUHSOHLEN	SUB		133
14-n	SCHULDSCHEIN	SUB		66
10	SENNHIRT	SUB		148
11	SOMMERSCHLUSSVERKAUF	SUB		86
11	SONNENBRAND	SUB		194

10	SPINNT	VRB		170
5-es	STAATSMANNES	SUB		189
10	STAMPFT	VRB		174
14-e	STEINE	SUB		136

<S. 139>

HO	*Wortform*	*richtig*	*falsch*	*Satznummer*
8	STOLZIERT	VRB		258
14-e	TAPETE	SUB		7
14-en	TATAREN	(SUB)	ADV	192
„	„	SUB		166
14-n	TERPENTIN	SUB		261
14-e	TRAUTE	VRB		102
14-en	TULPEN	SUB		252
10	UEBERGEORDNET	PTZ		170
14-en	UEBERHOEREN	INF		11
14-e	UEBERREICHTE	VRB		211
10	UEBERVORTEILT	PTZ		196
14-e	UEBERWINDE	VRB		234
14-en	UNBEKLEIDETEN	ADJ		169
14-e	UNGEAHNTE	ADJ		206
14-en	UNGEWASCHEN	(ADV)	SUB	174
55	UNI	SUB		34
10	UNTERHIELT	VRB		247
14-e	UNTERHOSE	SUB		167
14-en	UNTERHOSEN	(SUB)	ADV	170
14-e	UNUEBLICHE	ADJ		134
14-en	VEGETATIONSARMEN	ADJ		115
14-en	VERBISSEN	(ADV)	SUB	174
10	VERBRAUCHERZEITSCHRIFT	(SUB)	PTZ	207
14-e	VERHAUE	VRB		126
10	VERKLEBT	VRB		1
10	VERSCHMITZT	(ADV)	SUB	87
14-en	VERZEIHEN	INF		170
11	VETTEL	SUB		180
14-en	VOLKSWAREN	SUB		206
„	„	SUB		236
11	VORFALL	SUB		249
11	VORNEHM	ADV		281
14-en	WALDANHOEHEN	SUB		183
14-e	WANGE	SUB		14
14-n	WARENHAESERN	SUB		251
14-e	WELTFRIEDE	SUB		110
14-e	WERKZEUGE	SUB		143
„	“	SUB		200
14-e	WESTE	SUB		151
14-n	WETTERHAHN	SUB		19
11	WINTERSCHLUSSVERKAUF	SUB		251
11	WOCHENLANG	ADV		96
14-en	WOLKENFETZEN	(SUB)	ADJ	219
14-en	WUNDERSCHOENEN	ADJ		248

14-en	WUNDERTEN	(VRB)	ADV	23
20	ZAEHFLUESSIGER	ADJ		1
14-en	ZAEHLTEN	VRB		166
10	ZAHNARZT	SUB		50
55	ZEBRA	SUB		164
10	ZIELORT	SUB		194
55	ZOO	SUB		70
10	ZURUECKGEBRACHT	PTZ		60
14-en	ZURUECKZUERHALTEN	(INF)	VRB	9

<S. 140>

TABELLE VI Lösungsstatistik nach Wortklassen

Wortklasse	Mögliche Lösungen A	davon nicht erkannt B	A1	A2	B1	B2
SUB	245	45	116	129	23	22
VRB	83	12	63	20	7	5
ADJ	90	11	65	25	7	4
ADV	44	22	33	11	15	-7
PTZ	25	3	13	12	1	2
INF	12	8	4	8	3	5
Summe:	499	101	293	205	56	45

TABELLE VII Fehlertypen nach Wortklassen

richtig/falsch	*Anzahl*	*richtig/falsch*	*Anzahl*	*richtig/falsch*	*Anzhl*
ADJ / SUB	9	VRB / SUB	8	ADV / SUB	15
ADJ / VRB	2	VRB / ADJ	1	ADV / VRB	7
SUB / ADV	18	VRB / ADV	2		
SUB / VRB	12	VRB / PTZ	1	PTZ / SUB	1
SUB / PTZ	3	INF / SUB	3	PTZ / VRB	1
SUB / ADJ	12	INF / VRB	5	PTZ / ADV	1

TABELLE VIII Falsche Reduktionsergebnisse nach Wortklassen

Red.=	SUB	36
	VRB	27
	INF	-
	PTZ	4
	ADJ	13
	ADV	21
		101

<S. 141>

TABELLE IX Richtige Lösung der falsch gelösten Homographen nach Wortklassen

SUB	45

VRB	12
ADJ	11
ADV	22
INF	8
PTZ	3
	101

TABELLE X Lösungsstatistik nach Homographentypen I: Vorgegebene unbekannte Wortformen

HO	richtig	falsch	richtig						falsch					
			SUB	VRB	ADJ	PTZ	INF	ADV	SUB	VRB	ADJ	PTZ	INF	ADV
5-em	22	1	13	n	9	n	n	n	1	n	-	n	n	n
5-es	19	3	12	n	7	n	n	n	2	n	1	n	n	n
10-st	34	10	10	11	n	10	n	3	5	4	n	-	n	1
10-t	29	5	9	18	n	2	n	-	1	.	n	2	n	2
11	31	3	21	n	n	n	n	10	2	n	n	n	n	1
14-e	30	4	10	11	9	n	n	-	-	2	-	n	n	2
14-en	32	14	4	11	16	n	1	-	5	6	2	n	-	1
14-n	14	9	6	5	n	n	-	3	2	4	(1)	n	-	2
46	27	6	8	n	17	n	n	2	4	n	1	n	n	1
	238	55	93	56	58	12	1	18	22	16	5	2	-	10

TABELLE XI Lösungsstatistik nach Homographentypen II: nicht vorgegeben Wortformen

HO	richtig	falsch	richtig						falsch					
			SUB	VRB	ADJ	PTZ	INF	ADV	SUB	VRB	ADJ	PTZ	INF	ADV
4	1	-	-	-	n	n	1	n	-	-	n	n	-	n
5-em	5	2	5	n	n	-	n	n	2	n	-	n	n	n
5-es	6	-	3	n	3	n	n	n	-	n	-	n	n	n
8	1	-	n	1	n	-	n	-	n	-	n		n	
10-st	-	1	-	-	n	-	n	-	n	-	n	-	n	-
10-t	23	5	6	7	n	10	n	-	2	1	n	2	n	-
11	27	3	23	n	n	n	n	4	2	n	n	n	n	1
14-e	38	8	27	5	6	n	n	-	1	2	2	n	n	3
14-en	33	18	19	2	10	n	2	-	4	7	3	n	-	4
<S. 142>														
14-n	10	3	10	-	n	n	-	-	2	-	n	n	-	1
20	1	-	n	n	1	n	n	-	n	n	-	n	n	–
46	9	5	8	n	1	n	n	-	-	n	3	n	n	2
55	6	1	6	n	n	n	n	n	1	n	n	n	n	n
	160	46	107	15	21	10	3	4	14	11	8	2	-	11

TABELLE XII: Lösungsstatistik nach Homographentypen III: Gesamtstatistik der unbekannten Wortformen

HO	richtig	falsch	richtig						falsch					
			SUB	VRB	ADJ	PTZ	INF	ADV	SUB	VRB	ADJ	PTZ	INF	ADV

4,8,20														
	3	-		1	1		1							
5-em	27	3	18	n	9	n	n	n	3	n	-	n	n	n
5-es	25	3	15	n	10	n	n	n	2	n	1	n	n	n
10-st	34	11	10	11	n	10	n	3	5	5	n	-	n	1
10-t	52	10	15	25	n	12	n	-	3	1	n	4	n	2
11	58	6	44	n	n	n	n	14	4	n	n	n	n	2
14-e	68	12	37	16	15	n	n	-	1	4	2	n	n	5
14-en	65	32	23	13	26	n	3	-	9	13	5	n	-	5
14-n	24	12	16	5	n	n	-	3	4	4	(1)	n	-	3
46	36	11	16	n	18	n	n	2	4	n	4	n	n	3
55	6	1	6	n	n	n	n	n	1	n	n	n	n	n
	398	101	200	71	79	22	4	22	36	27	13	4	-	21

<S. 143>

TABELLE XIII Aufgliederung der falschen Lösungen nach möglichen richtigen Homographenlösungen

HO	S	A					B						
		SUB	VRB	ADJ	PTZ	INF	ADV	SUB	VRB	ADJ	PTZ	INF	ADV
5-em	3	-	n	1	n	n	n	-	n	2	n	n	n
5-es	3	1	n	2	n	n	n	-	n	-	n	n	n
10-st	11	2	-	n	1	n	7	1	-	n	-	n	–
10-t	10	2	n	-	n	-	2	1	n	-	n	2	
11	6	1 n	n	n	n	2	(1)	n	n	n	n	2	
14-e	12	4 -	-	n	n	-	6	1	1	n	n	-	
14-en	32	7	4	2	n	1	-	7	2	1	(2)	4	2
14-n	12	3 -	n	n	2	4	1	1	n	n	1	-	
46	11	2 n	2	n	n	2	5	n	-	n	n	-	
55	1	-	n	n	n	n	n	-	n	n	n	n	(1)
	101	22	7	7	1	3	15	23	5	4	2	5	7

HO	S	gesamt					
		SUB	VRB	ADJ	PTZ	INF	ADV
5-em	3	-	n	3	n	n	n
5-es	3	1	n	2	n	n	n
10-st	11	3	-	n	1	n	7
10-t	10	4	4	n	-	n	2
11	6	2	n	n	n	n	4
14-e	12	10	1	1	n	n	–
14-en	32	14	6	3	(2)	5	2
14-n	12	4	1	n	n	3	4
46	11	7	n	2	n	n	2
55	1	-	n	n	n	n	(1)
	101	45	12	11	3	8	22

<S. 144>

TABELLE XIV Satzlängenstatistik

Anzahl je Satz	*Sätze*	*Wörter insges.*	*Strichliste*
1	-	-	
2	1	2	I
3	6	18	IIIIII
4	14	56	IIIIIIIIIIIII
5	18	90	IIIIIIIIIIIIIIIII
6	17	102	IIIIIIIIIIIIIIII
7	24	168	IIIIIIIIIIIIIIIIIIIII
8	36	288	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
9	17	153	IIIIIIIIIIIIIII
10	23	230	IIIIIIIIIIIIIIIIIII
11	22	242	IIIIIIIIIIIIIIIIII
12	13	156	IIIIIIIIIIII
13	9	117	IIIIIIII
14	13	182	IIIIIIIIIIII
15	15	225	IIIIIIIIIIIII
16	7	112	IIIIII1
17	8	136	IIIIII11
18	6	108	IIIIII
19	3	57	III
20	7	140	IIIIII1
21	3	63	III
22	1	22	I
23	5	115	IIIII
24	-	-	
25	3	75	III
26	3	78	III
27	2	54	II
28	-	-	
29	1	29	I
30	2	60	II
31	-	-	
32	1	32	I
33	1	33	I
34	-	-	
35	1	35	I
	282	3178	

<S. 145>

TABELLE XV Gesamtstatistik zur Analyse der Zeitungsartikel

SATZ	*WORTFORM*	*ERG.WK 145/1 H0145/3 RICHT.WK 145/4*	*MORPH. 145/2*
556	100-JAEHRIGE	ADJ	vor 1.W. 145/5
541	1000	ADJ	"
548	120000	ADJ	"
530	15	SUB	"
532	15	SUB	"

517	1918	SUB	"		
547	1966-1967	SUB	"		
500	1971	SUB	"		
550	1971	SUB	"		
504	1975	SUB	"		
524	25	ADJ	"		
548	300000	ADJ	"		
526	5	SUB	"		
541	500	SUB	"		
547	500	ADJ	„		
529	ABENTEUERLICHEN	ADJ	LICH		
510	AFRIKA	SUB	A	55	
513	AFRIKA	ADV	A	55	SUB
521	AFRIKA	SUB	A	55	
512	AKTIVIERUNG	SUB	UNG		
505	ALS-OB-TARIFE	SUB	E	14	
511	AMTSSPRACHE	VRB	E	14	SUB
500	ANHEBUNG	SUB	UNG		
502	ANSCHLUSSES	SUB	ES	5	
519	ARABISCH	ADV	ISCH		
542	AUFFANGBEHAELTER		SUB	ER	46
531	AUFGEBAUT	PTZ	T	10	
544	AUFGEHOBEN	VRB	EN	14	PTZ
517	AUFWAERTSENTWICK-LUNG	SUB	UNG		
500	AUSGEBLIEBENE	VRB	E	14	ADJ
500	AUSNAHMETARIFE	SUB	E	14	
502	AUSNAHMETARIFE	SUB	E	14	
503	AUSNAHT TARIFE	SUB	E	14	
504	AUSNAHMETARIFE	SUB	E	14	
500	ANSAESSIGEN	ADJ	IG		
508	AUSWEISUNG	SUB	UNG		
537	AUTOMAT	SUB	T	10	
544	AUTOMAT	PTZ	T	10	SUB
535	AUTOMATEN	ADV	EN	14	SUB
542	AUTOMATEN	SUB	EN	14	<S. 146>
545	BAHAMAS	ADV	S	11	SUB
529	BANANARIVER	SUB	ER	46	
526	BERSTENDE	ADJ	END		
548	BESTEUERT	PTZ	T	10	
545	BESTEUERTEN	ADJ	EN	14	
515	BILDUNGSPLANS	SUB	S	11	
508	BUERGERRECHTSOR-GANISATIONEN	SUB	ION		
554	COMPUTER	SUB	ER	46	
521	DEUTSCHLEHRER	SUB	ER	46	
520	DEUTSCHSCHUELER	SUB	ER	46	
514	DEUTSCHUNTERRICHT		SUB	T	10
518	DEUTSCHUNTERRICHT		SUB	T	10
538	DREHENDE	ADJ	END		
531	EHRENGAESTE	SUB	E	14	
538	EINFUELLSTUTZEN	SUB	EN	14	
529	EINGRENZEN	SUB	ENZ		VRB

516	ELFENBEINKUESTE	ADV	E	14	SUB
515	ENGLISCH	ADV	ISCH		
535	ENTSAFTET	PTZ	T	10	
537	ENTSAFTEN	INF	EN	14	
511	ENTSENDEN	ADJ	END		INF
517	ENTWICKLUNGSPRO-JEKTEN	SUB	EN	14	
526	ERDRUECKT	VRB	T	10	
547	FINANZAEMTER	SUB	ER	46	
521	FRANCOSPANISCHEN	SUB	ISCH		ADJ
510	FRANKOPHONE	ADJ	E	14	
513	FRANKOPHONEN	SUB	EN	14	ADJ
521	FRANKOPHONEN	SUB	EN	14	ADJ
512	FREUNDSCHAFTSVER-TRAG	SUB	G	11	
539	FRUCHTFLEISCH	ADV	ISCH	SUB	
525	FUENF	ADJ	(ZAHL)		
543	FUENF	ADJ	(ZAHL)		
513	FUENFZIG	ADJ	"		
521	FUENFZIG	ADJ	"		
556	FUNKTIONSGERECH-TES	ADJ	ES	5	
536	GERAETETYPEN	VRB	EN	14	SUB
542	GESCHLEUDERT	PTZ	T	10	
532	GLEISSEND	ADV	END		
528	HERANROLLT	SUB	T	10	VRB
529	INDIANRIVER	SUB	ER	46	
525	KENNEDY-RAUM-FLUGZENTRUM	SUB	M	11	
543	KILO	SUB	O	55	
522	KLIRRT	ADV	T	10	VRB
555	KNOPFDRUCK	SUB	K	11	
517	KOLONIALMACHT	PTZ	T	10	SUB
517	KOLONIEN	SUB	IEN		
519	KONKURRANZKAMPF	HUB	F	11	
510	LEHRKRAEFTE	SUB	E	14	
513	LEHRKRAEFTE	SUB	E	14	
552	LEUCHTDICHTE	SUB	E	14	
554	LEUCHTDICHTE-BERECHNUNGEN	SUB	UNG		
552	LEUCHTDICHTEGLEICH-MAESSI(GKEIT)	SUB	I	55	
552	LICHTINGENIEURE	ADV	E	14	SUB
552	LICHTTECHNISCHE	ADJ	ISCH		
545	LICHTENSTEIN	VRB	N	14	SUB
516	LOGIKSCHULUNG	SUB	UNG		
524	LUFTLINIE	SUB	IE		<S. 147>
529	LUXUSLIMOUSINEN	SUB	EN	14	
553	MESSDATEN	VRB	EN	14	SUB
531	IVIETERBREITE	VRB	E	14	ADJ
507	MEXIKO	SUB	O	55	
547	MILLIONAERE	SUB	E	14	
529	MONDFLUGHAFEN	VRB	EN	14	SUB

531	MORGENSICHT	SUB	T	10		
524	MOTELS	SUB	S	11		
531	NACHRICHTEN-LABORS	SUB	S	11		
503	NICHTBERUECKSICH-TIGUNG	SUB	UNG			
553	NORMMASSTAB	SUB	B	11		
540	OEFFNUNGEN	SUB	UNG			
506	PERSONALAUSWEISE	SUB	WEISE			
507	PERSONALAUSWEISE	SUB	WEISE			
503	PRAEFERENZSPANNE	SUB	E	14		
552	QS-VERFAHREN	SUB	EN	14		
528	RAMPE	SUB	E	14		
538	RASPELSCHEIBE	SUB	E	14		
548	REGIERUNGSENTWURF		SUB	F	11	
511	REGIERUNGSSCHULEN		SUB	EN	14	
543	RUECKSTAENDE	SUB	END			
550	RUECKWIRKEND	ADV	END			
504	SAARWIRTSCHAFT	ADV	HAFT		SUB	
536	SAFT	SUB	T	10		
540	SAFT	SUB	T	10		
544	SAFT	SUB	T	10		
535	SAFTAUTOMAT	SUB	T	10		
544	SAFTTRINKER	SUB	ER	46		
535	SAFTZENTRIFUGE	SUB	E	14		
526	SATURN	ADV	N	14	SUB	
524	SATURN-5-TRAEGER-RAKETE	SUB	E	14		
502	SCHIFFAHRTSWEGE	SUB	E	14		
531	SCHLACHTENBUMM-LER	SUB	ER	46		
518	SCHUELERANDRANG	SUB	G	11		
514	SENEGAL	SUB	AL		NAM	
516	SENEGAL	SUB	AL		NAM	
539	SIEBKORB	SUB	(ZAHL)			
542	SIEBKORB	ADJ	(ZAHL)		SUB	
540	SIEBKORBES	ADJ	(ZAHL)		SUB	
552	SIEMENS	SUB	S	11		
556	SIEMENS-LICHTINGE-NIEURE	ADJ	E	14	SUB	
529	SONNTAGABEND	ADV	END	SUB		
526	STARTENDEN	SUB	END	ADJ		
524	STARTRAMPE	HUB	E	14		
525	STARTRAMPE	SUB	E	14		
523	STARTS	SUB	S	11		
549	STEUERFLUECHTLING		SUB	S	11	
546	STEUERFLUCHT	SUB	T	10		
545	STEUERLAST	SUB	T	10		
545	STEUEROASEN	SUB	EN	14		
547	STEUEROASEN	SUB	EN	14		
550	STEUEROASENGESETZ		NAM	Z	SUB	<S. 148>
549	STEUERPFLICHTIGEN	ADJ	IG			
549	STEUERSATZES	SUB	ES	5		

553	STRASSENBELEUCH-TUNG	SUB	UNG		
529	STRASSENBOESCHUN-GEN	SUB	UNG		
552	STRASSENDECKE	SUB	E	14	
534	TAEGLICHE	ADJ	E	14	
500	TARIFERHOEHUNG	SUB	UNG		
503	TARIFERHOEHUNG	SUB	UNG		
530	TELEKAMERAS	SUB	S	11	
530	TELESKOPE	SUB	E	14	
539	TRESTER	SUB	ER	46	
542	TRESTER	SUB	ER	46	
531	TROPISCHHEISSEN	ADJ	EN	14	
530	TURMARTIGE	ADJ	IG		
529	VEHIKELN	SUB	N	14	
533	VEREBBTE	VRB	E	14	
516	VERSUCHSGYMNA-SIEN	SUB	IEN		
505	VIERJAHRESPERIODE	SUB	(ZAHL)		
544	VITAMINAERMERE	ADJ	E	14	
552	VORAUSBERECHNUNG		SUB	UNG	
517	VORAUSGESEHEN	VRB	EN	14	PTZ
535	VORWEG	SUB	G	11	ADV
519	WAHLPFLICHTFACH	ADV	H	11	SUB
549	WELTEINKOMMEN	SUB	EN	14	
502	WETTBEWERBSTEL-LUNG	SUB	UNG		
528	ZAUNGAESTEN	SUB	EN	14	
514	ZEHN	ADJ	(ZAHL)		
545	ZEHN	ADJ	(ZAHL)		
546	ZEHNJAHRESFRIST	SUB	(ZAHL)		
518	ZENTRALAFRIKANI-SCHEN	ADJ	ISCH		
540	ZENTRIFUGALKRAFT	SUB	T	10	
537	ZENTRIFUGE	SUB	E	14	
539	ZENTRIFUGE	SUB	E	14	
538	ZERKLEINERT	PTZ	T	10	
538	ZERTEILTE	ADJ	E	14	
556	ZUKUNFTSWEISENDE	ADJ	END		
556	ZUKUNFTSFOR-SCHUNG	SUB	UNG		
517	ZWEI	ADJ	(ZAHL)		<S. 149>

PROBE EINES FERNSCHREIBERPROTOKOLLS DER LOCHSTREIFEN FÜR DIE TEXTEINGABE

Die Probe umfaßt die Testsätze der 3. Testperson (Satznummern 54 - 79). Die Großschreibung eines Anfangsbuchstabens ist nicht markiert, d.h. diese Information steht auch für die Analyse nicht zur Verfügung. Das Zeichen >< steht für ein Fragezeichen.

WARUM WIRD DAS SONNENLICHT IM LAUFE DER JAHRMILLIONEN NICHT GELBER>< BRANDT IST BUNDESPOLITIKER. IN AMERIKA UND

EUROPA BEDIENT MAN SICH GANZ ANALOGER COMPUTER. WER TRINKT NICHT LIEBER DUNKELBRAUNES ALS HELLES BIER>< WENN DIE AUSMASSE DIESES MACHTKOMPLEXES SICHTBAR WERDEN, WIRD DER WELTFRIEDE GEFAEHRDET SEIN. MAN BEHAUPTET OFT, DER MARS WAERE VON KLEINEN DUNKELGRUENEN MAENNCHEN BEWOHNT. DIESE LEUTE BEANSTANDETEN ABER SEHR DAS FEHLEN VON SAUERSTOFF, WENN SIE DA LEBTEN. JEDENFALLS IST GANZ SICHER, DASS ES KEINE KOSAKEN AUF DEM MARS GIBT. DA ES NUR WENIG WASSER GIBT, BEDECKEN WOHL NUR KLEINE FLECHTEN DEN KARGEN BODEN. DUNKELBRAUN LEUCHTEN DIE VEGETATIONSARMEN GEBIETE IM SOMMER, WAEHRENDSIE IM WINTER EINE HELLERE TOENUNG ZEIGEN. WIE MAG ES IN DIESER EINSAMKEIT AUSSCHAUEN, WENN EINEN DIE BEIDEN MONDE PHOBOS UND DEIMOS ANLAECHELN>< SICHER GRINST PHOBOS FRECH, WEIL ER NOCH SO KLEIN IST UND SICH NICHT BENEHMEN KANN. HAST DU DEIN VERSPRECHEN EINGELOEST, UNS ETWAS UEBER DIE KRANKHEITSERREGER AUF DEM MARS ZU BERICHTEN>< JA, DIE FLECHTEN WERDEN ERSTAUNLICH OFT VON EINER ART VON PEST BEFALLEN. SIE WACHSEN DANN FELSENFEST ZUSAMMEN. DAS IST EIN GEWAECHS, DAS NUR MIT AXT UND BEIL ZERSCHLAGEN WERDEN KANN. DA STAUNT IHR WOHL, DA BEGAFFT IHR MICH. KRAXELT NICHT AUF DEM SOFA HERUM, WENN ICH EUCH ETWAS ERZAEHLE. MAN SIEHT ZWAR DUNKELBRAUNEM STOFF NICHT AN, WIE SEHR IHR IHN STRAPAZIERT, ABER HOERT JETZT ENDLICH AUF. WER HAT DIESES UNANSTAENDIGE PHONEM VON SICH GEGEBEN>< ES BEDARF NUR EINES KLEINEN ANLAUFS, BIS ICH RICHTIG IN FAHRT KOMME UND EUCH VERHAUE. WIR GINGEN LANGSAM BERGAB, ALS WIR AUF EINE INTERESSANTE SPUR STIESSEN. ES HANDELTE SICH ZWEIFELLOS UM DEN ABDRUCK EINER PFOTE. DER DUNKELBRAUNE BODEN LIESS UNS DIE FAEHRTE NUR SCHWER ERKENNEN. SCHLIESSLICH WAREN WIR UNS EINIG, DASS NUR EIN TIER, NAEMLICH EIN ANGSTHASE, SOLCHE SPUREN HINTERLASSEN KANN. ICH VERBEUGE MICH VOR DEM ALLWISSENDEN COMPUTER. <S. 150>

SCHNELLDRUCKERPROTOKOLL VON ANALYSIERTEN TESTSÄTZEN OHNE KURZZEITLEXIKON (AUSWAHL)

Erklärung der wichtigsten Symbole[150/1]:

Spalte	Inhalt	Zeichen	Bedeutung
SZ	Satzzeichen (nach dem Wort)	K	Komma
		P	Punkt
		A	Ausrufezeichen
		F	Fragezeichen
		G	Gedankenstrich
HO	Homographenklasse		
WA	Ergebniswortklasse[150/2]		
D	Homographenprogramm-durchlaufzähler		
NO-V	Nominale oder Verbale Gruppe	NOM	NOG[150/3] im Nominativ
		AKK	NOG im Akkusativ

		DAT	NOG im Dativ
		PRF	Präpositionalgruppe
		ADB	Adverb/Prädikativum auf Gruppenebene
		VEG	Verbalgruppe(n)(teil)
TYP	Satztyp	HS	Hauptsatz
		GTS	Gliedteilsatz
AE	Analyseeinheit	n	Nr. der AE
UNT-GR bzw. END-GR	Subsatz- oder AE-Typ		
		HAUPTS	Hauptsatz
		ADVERS	Adverbialsatz
		RELATS	Relativsatz
		NEBENS	Objekt/Subjektsatz
		FRAGES	Fragesatz <S. 151>
WB	Bezugswerte	bei WA = REL	Bezugswort-Nr.
		bei WA = VRB	Rektionsangabe
BE	Subgruppen-Hinweis	AB bei VEG	Diskont. VEG 2. Teil
		PE	Präpositionale Subgr.
		AN	Anreihung (nominal)
		AR	Anreihung (Subsatz)
		AT	Attribut
RE, LE	Verweisnummern auf zusammengefasste AE	n unter RE	Zusammenfassung mit einer folgenden AE
		n unter LE	Zusammenfassung mit einer vorhergehenden AE
VG	Verbalgruppentyp[151/1]	n	Nr. des Typs
SOND	Zusatzangaben (Auswahl)	REV	Reflexiver Gebrauch möglich
		PAS	Passiv möglich
		ZZI	"ZU" zum Infinitiv
		NKO	Nebenordnende Konjunktion
		MOR	Wort morphologisch analysiert
		LWM	Wort ist in Kurzzeitlexikon aufzunehmen.

Weitere Hinweise bei der Erläuterung der Satzbeispiele

<S. 152>

SATZ 20006 WORTZAHL 14 ART

N	WORTLAUT	SZ	HO	WA	D	NO-V	TYP	AE	UNT-GR	END-GR	WB	BE	RE	LE	MK	VG	S	SOND
1	ES			PER	1	NOM	MS	1	MEHN	2 HAUPTS						1	1	BIN NOD
2	SAH			VRB	1	VEG					12							ACI IMI IMK
3	AUS		16	VZS	2													VEZ
4	WIE		26	PRP	1	PRF												
E	DUNKELBRAUNES NAH		6	ADJ	1													MAR LWM MOR
6	GROBES			ADJ	1													GST
7	KAFFEEPULVER NAH	K	46	SUB	1													MAR LWM MOR
8	UND		6	KON	1	NKO	EKO	MS		2	KONVRB		HAUPTS					
9	ER			PER	1	NOM												BIN NOD

10	HATTE			VRB 1 VEG													INK IMI HAB
11	NOCH		48	ADV 1 ADB													ROA UKZ
12	NIE			ADV 1 ADB													
13	DAVON		53	ADV 1 ADB													VEZ
14	GERAUCHT MAR,LWMIMOR	P	10	PTZ 1 VEG							12	AB					PMB PAS

Erläuterung: Drei Wortformen des Satzes konnten weder über das Lexikon noch morphologisch klassifiziert werden und wurden als "künstliche" Homographen (DUNKELBRAUNES HO 5 = ADJ/SUB; KAFFEEPULVER HO 46 = ADJ/SUB/ADV; GERAUCHT HO 10 = ADV/PTZ/VRB/SUB) gekennzeichnet. "Natürlich mehrdeutig" waren weiterhin AUS, WIE, NOCH und DAVON. Bis auf AUS wurden alle im 1. Durchlauf (richtig) gelöst, AUS daraufhin im 2. Durchlauf (D = 2). als Verbzusatz ermittelt. Die Informationen unter den Rubriken NO-V bis ENDGR zeigen den Beginn einer entsprechenden Gruppe an: Die Präpositionalgruppe PRF besteht also aus „WIE DUNKELBRAUNES GROBES KAFFEEPULVER"; der erste Hauptsatz beginnt bei Wort N 1, der zweite Hauptsatz bei Wort N 8.

<S. 153>

SATZ 20012 WORTZAHL 23

N	WORTLAUT	SZ	HO	WA	D	NO-V	TYP	AE	UNT-GR	END-GR	WB	BE	RE	LE	MK	VG	S	SOND
1	SEINEN			POS 1		AKK	MS	1	MEHN 1	HAUPTS						1	1	
2	EINDRUCK			SUB 1														
3	VOM			PRP 1								PE						PUA
4	SPIEGELSAAL			SUB 1														MOR
6	BESCHRIEB			VRB 1		VEG					12							IMI
6	DER		43	DEM 1		NOM												
7	STAATSMANN			SUB 1														
8	MIT		15	PRP 6		PRF												VEZ
9	DEN		43	DEM 6														
10	WORTEN	K		SUB 1														
11	ER			PER 1		NOM	MS	2	MEHN 1	HAUPTS						6	1	SIN NOD
12	HABE		4	VRB 1		VEG												INK HAB
13	DAS		43	DEM 1		AKK												
14	GEFUEHL			SUB 1														
16	GEHABT	K	8	PTZ 1		VEG					8	AB						REV PHB PAS
16	DASS			KON 1		EKO	3		NEBENS	1 2 KOS								
17	IHN			PER 1		AKK												
18	VON			PRP 1		PRF												
19	ALLEN			IND 1														
20	SEITEN			SUB 1														
21	BERUEHMTE			ADJ 1		NOM												GST
22	STAATSMAENNER			SUB 1														
23	ANLAECHELN NAM	P	14	VRB 1		VEG					12							PUA LWM MOR

Erläuterung: Zwei Wörter waren nicht im Lexikon: SPIEGELSAAL und ANLAECHELN. SPIEGELSAAL war zunächst (nicht. erkennbar) fälschlich "morphologisch" - Suffix -AL - (siehe SOND = MOR) als ADJ klassifiziert worden und wurde über die eingefügte Substantivierungsregel schließlich als SUB bestimmt. Die Wortfolge VOM SPIEGELSAAL konnte damit als präpositionale Subgruppe (PE) der ersten Nominalgruppe im AKK ermittelt werden. Da ANLAECHELN trotz der formalen "künstlichen" Mehrdeutigkeit ADV/VRB/SUB/INF richtig als VRB bestimmt werden konnte, war die Erkennung des Nebensatzes möglich. Neben diesen unbekannten Wörtern waren noch DER (N 6), DEN (N 9), HABE (N 12), DAS (N 13) und GEHABT (N 15, hier richtig PTZ, als VRB in GEHABT EUCH WOHL) über die Homographenauflösung zu klassifizieren.

<S. 154>

SATZ 20040 WORTZAHL 5

N	WORTLAUT	SZ	HO	WA	D	NO-V	TYP	AE	UNT-GR	END-GR	WB	BE	RE	LE	MK	VG	S	SOND

```
1 EINGELOEST          10  PTZ 1 VEG   MS 1   MEHN 2 HAUPTS  12              46  3 PSN PHB PAS LWMIFOR
2 WURDE                   VRB 1                                                    INK ALL IMI WRD
3 SEIN                36  POS 1 NOM                                              SN
4 VERSPRECHEN          4  SUB 1                                                   REV
5 NIE                 A   ADV 1 ADB
```

Erläuterung: Von fünf unbekannten Wörtern dieses Satzes sind drei syntaktisch mehrdeutig. Das unbekannte Wort EINGELOEST (HO 10, Mehrdeutigkeit SUB/PTZ/VRB/ADV konnte trotz der einigermaßen ungewöhnlichen Stellung richtig als PTZ klassifiziert werden; die weiteren natürlich mehrdeutigen Wörter SEIN (POS/SUB/INF) sowie VERSPRECHEN (SUB/VRB/INF) wurden ebenfalls richtig bestimmt.

<S. 155>

SATZ 20084 WORTZAHL 19

```
N WORTLAUT        SZ  HO  WA D NO-V  TYP AE  UNT-GR END-GR   WB BE  RE  LE  MK VG S SOND

1 ER                      PER 1    NOM  MS       1          VERB 1  HAUPTS       3
                      1   SIN NOD
2 HAT                     VRB 1 VEG                                               HAB
3 IHN             K       PER 1 AKK
4 ALS                 22  KON 1 EKO  GAD  2 ADVERS                          6   1 UKZ
6 HAETTE                  VRB 1 VEG                                               INK HAB
6 ER                      PER 1    NOM
                  SIN NOD
7 NIE                     ADV 1 ADB
8 SO                      48  ADV 1  ADB
                  KOA
9 ETWAS               14  ADJ 1                                  AT               PHB
10 SCHOENES               SUB 1 GST
11 GESEHEN        K   45  PTZ 1 VEG                          8   AB               PMB PAS
12 BEGAFFT            10  PTZ 1 VEG        3 VERB 8 HAUPTS  12  AB  1        6   1 PHB PAS MAR
    LWM FOR
13 DEN                43  DEM 1 AKK
14 GANZEN                 ADJ 1
16 TAG            K       SUB1
16 SOGAR                  ADV 1 ADB        4 PRAEAU PRAEAP  AR
17 MIT                15  PRP 2 PRF                                               VEZ
18 WACHSENDER         20  ADJ 1                                                   GST
19 BEGEISTERUNG   P       SUB 1
```

Erläuterung: Dieser recht komplexe Satz, bei dem darüberhinaus 8 von 19 Wörtern homograph sind, enthält ein unbekanntes Wort (BEGAFFT). Das Problem dabei war, dieses Wort als Teil der Verbalgruppe mit dem Wort HAT (N 2) zu verbinden, wobei der eingeschobene Adverbialsatz als solcher zu erkennen und entsprechend zu berücksichtigen war. Da auch die Auflösung der übrigen, „natürlich mehrdeutigen" Wortformen richtig war, konnte diese Einbettung des Adverbialsatzes ALS ... GESEHEN in den Hauptsatz ER HAT IHN / BEGAFFT DEN GANZEN TAG ... erkannt werden. Die Zahlenwerte 3 unter RE bzw. 1 unter LE zeigen beim ersten Wort der Hauptsatzteile diese Verbindung an. Die (ausgeklammerte) Nominalgruppe DEN GANZEN TAG ist (vorläufig) nur als AKK bestimmbar.

<S. 156>

```
N WORTLAUT        SZ  HO  WA D NO-V  TYP AE  UNT-GR END-GR   WB BE  RE  LE  MK VG S SOND

1   VERSCHMITZT       10  SUB 6 DAT   MS   1 MEHN 2 HAUPTS                  1   1 PSN PHB PAS
    MAR LWM
2   GRINST            10  VRB 6 VEG                                  12           PSN PHB PAS
    LWM FOR
3   FRITZ                 NAM 1 NOM
4   AUS               18  PRP 7 PRF                                               VEZ
5   DEM               43  DEM 7
6   FENSTER               SUB 1
```

```
7   HERAUS              53  ADV 6 ADB
8   UND                     KON 1 EKO  MS    2 KONVRB          HAUPTS     AR          1          1
                NKO
9   TUT                     VRB 1 VEG                          12                     REV
10  SONST                   ADV 1 ADB
11  SO              K   48  ADV 1 ADB                                                 KOA
12  ALS                 22  KON 1 EKO  GAD   3 ADVERS                        1    2  UKZ
13  WUESSTE                 VRB 1 VEG                          8                      IMR
14  ER                      PER 1 NOM  SIN NOD
15  VON                     PRP 1 PRF
16  NICHTS          P   14  SUB 1                                                     PHB
```

Erläuterung: In diesem Satz sind zwei unmittelbar aufeinander folgende Wortformen (VERSCHMITZT und GRINST) nicht anhand des Lexikons oder mithilfe der morphologischen Regeln zu klassifizieren gewesen. Beide wurden in Abhängigkeit voneinander im 6. Homographendurchlauf (D = 6) reduziert: Für VERSCHMITZT (HO 10 SUB/VRB/PTZ/ADV) wurde die formal-syntaktisch mögliche Lösung SUB gefunden. Hier zeigt sich die Schwäche des bisherigen Analyseverfahrens, nur eine Lösung anzubieten, wo weitere durchaus auf dieser sprachlichen Ebene möglich wären. Es zeigt sich aber auch, daß es notwendig ist, Methoden zu erarbeiten, die diesem Problem bei der automatischen Klassifizierung unbekannter Wörter Rechnung tragen.
Die hier gewählte Lösung wurde nach dem Strukturtyp "ANGST GRINST FRITZ AUS DEN AUGEN HERAUS" ermittelt.

<S. 157>

SATZ 20088 WORTZAHL 11

```
N WORTLAUT          SZ  HO  WA D NO-V  TYP AE  UNT-GR END-GR   WB BE  RE  LE  MK VG S SOND

1 NACH                  18  PRP 6 PRF   MS    1 VERB 1 HAUPTS                  10  1 VEZ
2 MONATELANGEN          14  ADJ 7                                                    INK LWM MOR
  NAM
3 REPARATURARBEITEN         14  SUB 8
                INK LWM MOR NAM
4 WAREN                 4   VRB 9 VEG                      1                        IMI SN
6 DIE                   43  DEM 1 NOM
6 BEANSTANDETEN         14  ADJ10                                                    INK LWM MOR
  NAM
7 MAENGEL           SUB 1
8 IMMER                 ADV 1 ADB
9 NOCH                  48  ADV 1 ADB                                            KOA UKZ
10                  NICHT       ADV 1 ADB
                    NEA
11                  BEHOBEN  P 8    ADV10 ADB
                        REV PHB PAS IMI
```

Erläuterung: Acht (!) von elf Wörtern dieses Satzes waren mehrdeutig; davon mußten die drei lexikalisch/morphologisch nicht klassifizierbaren über die Endgrapheme -EN der komplexesten Klasse HO 14 (mögliche Wortklassen: SUB/ VRB/INF/ADJ/ADV) zugeordnet werden. Zehn Durchläufe des Homographenreduktions-Programms waren nötig, ehe alle richtig bestimmt waren (Bei BEHOBEN zeigt die Angabe ADV an, daß es nicht Teil einer Verbalgruppe (PTZ) ist). Daß MONATELANGEN (D = 7), REPARATURARBEITEN (D = 8) und BEANSTANDETEN (D = 10) richtig klassifiziert werden konnten, ist sicherlich der einfachen Struktur des Satzes zuzuschreiben.

<S. 158>

SATZ 20104 WORTZAHL 19

N	WORTLAUT	SZ	HO	WA	D	NO-V	TYP	AE	UNT-GR	END-GR	WB	BE	RE	LE	MK	VG	S	SOND
1	MAN			PER	1	NOM	MS	1	MEMN 2	HAUPTS						1	3	SIN NOD
2	SAGT			VRB	1	VEG					12							REV
3	VON			PRP	1	PRF												
4	IHM	K		PER	1													
5	ER			PER	1	NOM	MS	2	MEHN 1	HAUPTS						1	1	SIN NOD
6	VERBEUGE		14	VRB	1	VEG					12							MAR INK LWM MOR NAM
7	SICH			PER	1	AKK												SIN PLU
6	VOR		15	PRP	1	PRF												VEZ
9	ALLEN	K		IND	1													
10	VON			PRP	1	PRF	GTS	3	RELATS							1	2	
11	DENEN		43	REL	1						9							
12	ER			PER	1	NOM												SIN NOD
13	MEINT	K		VRB	1	VEG	8											
14	DASS			KON	1	EKO	4		NEBENS	64 2 KOS								
16	SIE			PER	1	NOM												SIN PLU NOE
16	IHM			PER	1	DAT												
17	NUETZLICH			ADV	1	ADB												GST
18	SEIN		36	INF	1	VEG					8							SN
19	KOENNTEN	P		VRB	1													MDV IMK

Erläuterung: VERBEUGE (N 6, HO 14-e) war eines der vorgegebenen unbekannten Wörter. Dieser 19-Wort-Satz besteht immerhin aus vier Subsätzen; dennoch konnten die vier syntaktisch mehrdeutigen Wortformen alle im 1. Homographendurchlauf (wohl aufgrund des engeren Kontexts) richtig klassifiziert werden.

<S. 159>

SATZ 20128 WORTZAHL 8

N	WORTLAUT	SZ	HO	WA	D	NO-V	TYP	AE	UNT-GR	END-GR	WB	BE	RE	LE	MK	VG	S	SOND
1	WER		3	FRA	1	NOM	FS	1	FRAGES							6	1	INK
2	HAT			VRB	1	VEG												INK HAB
3	DIESES			DEM	1	AKK												
4	UNANSTAENDIGE			ADJ	1													MAR LWM MOR NAM
5	PHONEM		5	SUB	1													
6	VON			PRP	1	PRF												SIN PLU
7	SICH			PER	1													
8	GEGEBEN	F	45	PTZ	1	VEG					8	AB						PMB PAS

Erläuterung: Ein Beispiel für einen unkomplizierten Fragesatz: Das vorgegebene unbekannte Wort PHONEM (Endung -EM: ADJ/SUB/ADV als mögliche Wortklassen) ließ sich durch den Kontext leicht richtig bestimmen; auch die Verbalklammer HAT ... GEGEBEN bot keine Reduktionsprobleme.

<S. 160>

SATZ 20219 WORTZAHL 15 ART

N	WORTLAUT	SZ	HO	WA	D	NO-V	TYP	AE	UNT-GR	END-GR	WB	BE	RE	LE	MK	VG	S	SOND
1	VOM			PRP	1	PRF	MS	1	MEMN 2	HAUPTS						1	1	PUA
2	STURM			SUB	1													
3	GEPEITSCHTE		14	VRB	1	VEG					12							MAR INK LWM MOR KAM
4	WOLKENFETZEN		14	ADJ	6	NOM												INK LWM MOR NAH
5	BEDECKEN		14	SUB	6													INK LWM MOR NAM
6	DEN		43	DEM	1	AKK												
7	HIMMEL	K		SUB	1													
8	GRAU		11	SUB	1	NOM		2	NOM 1	APPOSI								GST
9	IN			PRP	1	PRF												
10	GRAU	K	11	SUB	1													GST
11	UNHEILVERHEISSEND			ADV	1	ADB	MS	3	MEMN 2	HAUPTS						1	1	MOR

```
12  ZIEHT                VRB 1 VEG                          12                REV
13  EIN              27  IND 1 NOM                                            VEZ
14  GEWITTER             SUB 1
16  AUF              P   VZS 1 VEG                              AB            VEZ
```

Erläuterung: Alle drei (aufeinanderfolgenden !) unbekannten Wortformen (GEPEITSCHTE WOLKENFETZEN BEDECKEN) wurden "falsch" aufgelöst: Berücksichtigt man jedoch nur die Stellung der Wortklassen, so erscheint diese Auflösung durchaus sinnvoll; sie entspräche etwa dem Satz DURCHS FERNGLAS BEOBACHTETEN KLEINE KINDER DEN HIMMEL ... . Eine feinere Homographenauflösung hätte hier bei Verwendung exakterer Informationen (Numeruskongruenz, Stellung des Adjektivs nach bestimmtem / unbestimmtem Artikel) durchaus diese Fehlanalyse vermeiden helfen.

<S. 161>

SATZ 20275 WORTZAHL 6

```
N WORTLAUT          SZ  HO  WA D NO-V  TYP AE  UNT-GR END-GR   WB BE  RE  LE  MK VG S SOND

1  PEST                 10  SUB 2 NOM  MS    1  MEMN 2 HAUPTS                 1  1 PSN PMB PAS
   LWM MOR
2  UND                      KON 1 AN                                           NKO
3  BUCA                 55  SUB 1                                              LWM MOR
4  LIEGEN               4   VR8 1 VEG                          4
6  EINANDER                 PER 1 AKK
6  GEGENLIEBER      P   44  VZS 1 VEG                              AB          VEZ
```

Erläuterung: In diesem kurzen Satz waren zwei Wörter "natürlich mehrdeutig" (LIEGEN und GEGENUEBER), zwei weitere befanden sich nicht im maschinellen Lexikon: Bei BUDA wurden (morphologisch-graphematisch) noch die Wortklassen SUB und NAM zugelassen, während PEST a priori noch SUB, (NAM), VRB, ADV und PTZ sein konnte. Da für die Auflösung der Mehrdeutigkeit SUB/NAM noch kein entsprechendes Programm erstellt ist, erklärt sich die Auflösung beider Wörter als SUB; der vom Autor des Satzes beabsichtigte Effekt (PEST - ein vorgegebenes Wort - als NAM) konnte also vom Lösungssystem her bereits nicht erreicht werden. Die natürlich mehrdeutigen Wortformen (besonders LIEGEN: zugleich Flexionsform des Substantivs DIE LIEGE) wurden entsprechend ihrer Funktion im Satz aufgelöst.

<S. 162>

SATZ 503 WORTZAHL 17

```
N WORTLAUT          SZ  HO  WA D NO-V  TYP AE  UNT-GR END-GR   WB BE  RE  LE  MK VG S SOND

1  DURCH                18  PRP 6 PRF  MS    1  MEHN 2 HAUPTS                 6  1 VEZ
2  DIE                  43  DEM 6
3  NICHTBERUECKSICHTIGUNG   SUB 1                                              MOK
4  DER                  43  DEM 1                              AT
5  AUSNAHMETARIFE       14  SUB 2                                              MAR INK LWM
   MOR NAM
6  DER                  43  DEM 1
7  LETZTEN                  ADJ 1                              AT              GST
8  TARIFERHOEHUNG           SUB 1                                              MOR
9  BEI                  16  PRP 7                                              VEZ
10                  DER      43  DEM 7                              AT
11                  BUNDESBAHN               SUB 1
12                  HABE         4  VRB 2
                    VEG INK HAB
13                  SICH             PER 1 AKK
                    SIN PLU
14                  DIE      43  DEM 1  NOM
15                  PRAEFERENZSPANNE          14        SUB 1
                                MAR INK LWM MOR 'NAM
16                  JEDOCH       48         ADV 1 ADB
                        KOA
17                  ERHOEHT      P  8       PTZ 2 VEG                          8   AB
                        REV PHB PAS
```

Erläuterung: Bei diesem Beispielsatz aus den Zeitungsartikel-Sätzen waren vier Wortformen nicht im maschinellen Lexikon enthalten, von denen zwei (NICHTBERUECKSICHTIGUNG, TARIFERHOEHUNG) über das Suffix -UNG in der morphologischen Liste eindeutig als Substantiv klassifiziert werden konnten. Die noch verbleibenden Wörter AUSNAHMETARIFE und PRAFERENZSPANNE wurden beide über die quasimorphologische Analyse (E als letztes Graphem) dem Homographentyp HO 14 (Mehrdeutigkeit SUB/VRB/ADJ/ADV) zugeordnet. In beiden Fällen wurde anhand des Kontexts die syntaktisch korrekte Information SUB erschlossen, so daß auch die durch eine Reihe von Attributen erweiterte Präpositionalgruppe N 1 - N 11 korrekt ermittelt werden konnte. Bei 12 der 17 Wortformen des Satzes mußte dabei die aktuelle Funktion durch die Homographenauflösungsprogramme ermittelt werden.

<S. 163>

SATZ 530 WORTZAHL 12

N	WORTLAUT	SZ	HO	WA	D	NO-V	TYP	AE	UNT-GR	END-GR	WB	BE	RE	LE	MK	VG	S	SOND
1	TELEKAMERAS		11	SUB	6	NOM	MS	1	VERB	1 HAUPTS						10	1	LWM NIW NAM
2	UND			KON	1							AN						NKO
3	TELESKOPE		14	SUB	7													MAR INK LWM
	MOR NAM																	
4	WAREN		4	VRB	7	VEG					1							IMI SN
5	AUF		18	PRP	8	PRF												VEZ
6	DAS		43	DEM	8													
7	FERNE		5	ADJ	1													GST
8	TURMARTIGE			ADJ	1													MOR
9	GEBILDE			SUB	1													
10	APOLLO			NAM	1								AT					
11	15			SUB	1													ZAL
12	GERICHTET	P	45	ADV	1	ADB												PHB PAS

Erläuterung: Ein Satz aus dem Zeitungsartikel-Test. Die Wortformen TELEKAMERAS sowie TELESKOPE waren noch über die Satzanalyse zu klassifizieren, da sie nicht im maschinellen Lexikon verzeichnet waren; TURMARTIGE wurde mittels des Suffix -IG morphologisch klassifiziert. APOLLO 15 wurde als Attribut (AT) zu GEBILDE geordnet: 15 war zunächst als Zahladjektiv (SOND = ZAL) ermittelt und später, da es nicht in der "normalen" attributiven Stellung stand, mittels des Substantivierungsprogramms als SUB klassifiziert worden. Auch in diesem Satz waren mehr als die Hälfte der Wortformen syntaktisch mehrdeutig.

<S. 164>

SATZ 539 WORTZAHL 18

N	WORTLAUT	SZ	HO	WA	D	NO-V	TYP	AE	UNT-GR	END-GR	WB	BE	RE	LE	MK	VG	S	SOND
1	IN			PRP	1	PRF	MS	1	VERB	1 HAUPTS			3				1	
2	DER		43	DEM	1													
3	ZENTRIFUGE		14	SÜB	1													MAR INK LWM
	MOR NAM																	
4	WIRD			VRB	1	VEG					1							INK WRD
5	DER		43	DEM	1	NOM												
6	TRESTER	G	46	SUß	1													MAR LWM MOR
	NAM																	
7	DAS		43	DEM	1	NOM	HS	2	VERB	1 HAUPTS						10	1	
8	SIND			VRB	1	VEG					1							SN
9	DIE		43	DEM	1	NOM												
10	RESTE			SUB	1													
11	VON			PRP	1	PRF												
12	SCHALEN		7	SUB	1													VEZ GST
13	UND			KON	1	EKO												NKO
14	FRUCHTFLEISCH	G		ADV	1	ADB												MOR
15	IN			PRP	1	PRF		3	VERB	8 HAUPTS				1		17	1	
16	EINEM			IND	1													
17	SIEBKORB			SUB	1													MOR ZAL NIW
18	GESAMMELT	P	45	PTZ	1	VEG						AB						PHB PAS

Erläuterung:
ZENTRIFUGE, TRESTER, FRUCHTFLEISCH, SIEBKORB - diese Wortformen des Zeitungsartikel-Satzes waren nicht im maschinellen Lexikon verzeichnet. Die beiden ersten wurden über die Homographenauflösungsprogramme (richtig) als Substantive erkannt, obwohl ZENTRIFUGE (präklassifiziert als HO 14 SUB/VRB/ADJ/ADV über das Endgraphem E) und TRESTER (präklassifiziert als HO 46 SUB/ADJ/ADV über die Endgrapheme ER) noch eine Reihe von Deutungen offenließen.
Die beiden übrigen Wortformen wurden mittels des morphologischen Erkennungsprogramms falsch klassifiziert: FRUCHTFLEISCH als ADV ("Suffix" ISCH) und SIEBKORB als Zahl-Adjektiv (ZAL unter der Rubrik SOND), da nur die ersten vier Grapheme - hier SIEB für SIEBEN - beim Zahlenerkennungsprogramm herangezogen werden. Über das Substantivierungsprogramm konnte allerdings SIEBKORB gleichsam wieder "korrigiert" werden: Ergebniswortklasse war SUB.
Der durch den in Gedankenstriche eingeschlossenen Hauptsatz N 7 - N 14 unterbrochene Satz N 1/N 6 ... N 15 / N17 wurde als zusammenhängend erkannt (vgl. die Informationen RE 3 und LE 1 beim ersten Wort der entsprechenden Analyseeinheiten).

<S. 165>

SATZ 547 WORTZAHL 24

N	WORTLAUT	SZ	HO	WA	D	NO-V	TYP	AE	UNT-GR	END-GR	WB	BE	RE	LE	MK	VG	S	SOND
1	DAMIT		48	ADV	6	ADB	MS	1	MEMN 2	HAUPTS						27	1	K0A
2	KOENNEN		2	VRS	7	VEG												ALL MDV
3	DIE		43	DEM	1	AKK												
4	FINANZAEMTER NAM		46	SUB	1													MAR LWM MOR
5	ZUMINDEST			ADV	1	ADB												
6	FUER			PRP	1	PRF												GST REV
7	EINIGE		29	IND	1													
8	JAHRE			SUB	1													ZSU
9	AUCH		48	ADV	1	ADB												NKD KOA WKZ
10	JENE			DEM	1	AKK												
11	RUND		14	ADJ	1													VEZ PHB
12	500			ADJ	1	ZAL												
13	MILLIONAERE NAM		14	SUB	8	MAR												INK LWM MOR
14	ERFASSEN	K	2	INF	8	VEG					8	AB						
15	DIE		43	REL	2	NOM	GTS	2		RELATS	13					43	2	
16	SICH			PER	1	AKK												SIN PLU
17	VOR		15	PRP	1	PRF												VEZ
18	ALLEM			IND	1													
19	SEIT		24	PRP	1	PRF												ZAL
20	1966-1967			SUB	1													
21	IN			PRP	1	PRF												
22	STEUEROASEN NAM		14	SUB	1													INK LWM MOR
23	ABGESETZT		45	PTZ	8	VEG												GST PHB PAS
24	HABEN	P	2	VRB	9						8				1			INK ALL HAB

Erläuterung: In diesem 24-Wort-Satz des Zeitungsartikel-Tests waren immerhin 15 Wörter syntaktisch mehrdeutig; drei davon (FINANZAEMTER, MILLIONAERE und STEUEROASEN) waren nicht im Lexikon vorhanden. Darüberhinaus mußten 500 und 1966-1967 als Zahlen ermittelt werden. MILLIONAERE, ERFASSEN und ABGESETZT wurden erst im 8. Homographendurchlauf klassifiziert; danach konnte erst das letzte Wort HABEN (richtig) als finites Verb bestimmt werden.

<S. 166>

ERGEBNISSE DES KURZZEITLEXIKONS (AUSWAHL)

Hinweis: Es handelt sich um die Ergebnisse der Wortklassenreduktion nach dem letzten der elf Analysezykleh. Alle 26 vorgegebenen unbekannten Wortformen sind in der Liste verzeichnet; dazu kommen einige nur einmal belegten Wortformen aus dem ersten Analysezyklus (etwa ALLSEITS, ASCHE), bei denen naturgemäß keine endgültige Reduktion möglich war.

Die nach den Wortklassensymbolen aufgeführten Zeichen bedeuten:

*	morphologisch nicht möglich
-	morphologisch zugelassen, doch aufgrund der Ergebnisse ausgeschlossen
8	Wortklasse endgültig bestimmt
1	Beleg für diese Wortklasse vorhanden
0	morphologisch möglich, doch bishernoch kein Beleg vorhanden

Die (nicht verzeichnete) Homographenklasse ergibt sich aus der Kombination der (noch) möglichen, also mit 8, 1 oder 0 gekennzeichneten Wortklassen.

<S. 167>

G.I. (ZI) LERNWOERTERBUCH SEITE 1 29.04.1970

ABDRUCK	ADJ*	ADV-	VRB*	INF*	PTZ*	SUB8	NAM-
ALLSEITS	ADJ*	ADV1	VRB*	INF*	PTZ*	SUB0	NAM0
ANALOGER	ADJ8	ADV-	VRB*	INF*	PTZ*	SUB1	NAM-
ANGSTHASE	ADJ0	ADV1	VRB1	INF*	PTZ*	SUB1	NAM0
ANLAECHELN	ADJ*	ADV1	VRB1	INF0	PTZ*	SUB1	NAM0
ANLAUFS	ADJ*	ADV-	VRB*	INF*	PTZ*	SUB8	NAM
ASCHE	ADJ0	ADV0	VRB0	INF*	PTZ*	SUB1	NAM0
AUSGERUHT	ADJ*	ADV0	VRB0	INF*	PTZ1	SUB0	NAM0
AXT	ADJ*	ADV1	VRB-	INF*	PTZ1	SUB8	NAM-
BEANSTANDETEN	ADJ1	ADV0	VRB1	INF0	PTZ*	SUB1	NAM0
BEDECKEN	ADJ1	ADV0	VRB1	INF1	PTZ*	SUB1	NAM0
BEGAFFT	ADJ*	ADV1	VRB1	INP*	PTZ1	SUB1	NAM0
BERGAB	ADJ*	ADV8	VRB*	INP*	PTZ*	SUB-	NAM-
BESONNEN	ADJ0	ADV1	VRB0	INF0	PTZ*	SUB0	NAM0
BESUCHERIN	ADJ*	ADV0	VRB0	INF0	PTZ*	SUB1	NAM0
BUNDESPOLITIKER	ADJ1	ADV1	VRB*	INP*	PTZ*	SUB8	NAM-
DIENSTJAHREN	ADJ0	ADV0	VRB0	INF0	PTZ*	SUB1	NAM0
DUNKELBRAUN	ADJ*	ADV1	VRB1	INF0	PTZ*	SUB1	NAM0
DUNKELBRAUNE	ADJ8	ADV-	VRB-	INP*	PTZ*	SUB1	NAM-
DUNKELBRAUNEM	ADJ8	ADV*	VRB*	INF*	PTZ*	SUB1	NAM-
DUNKELBRAUNES	ADJ1	ADV*	VRB*	INF*	PTZ*	SUB1	NAM0
DUNKELGRUENEN	ADJ8	ADV-	VRB-	INF-	PTZ*	SUB1	NAM-
EINGELOEST	ADJ*	ADV-	VRB-	INF*	PTZ8	SUB1	NAM-
MAMRSTUHL	ADJ*	ADV0	VRB*	INF*	PTZ*	SUB1	NAM0
FELSENFEST	ADJ*	ADV1	VRB1	INP*	PTZ1	SUB1	NAM0
FRISEURE	ADJ0	ADV0	VRB0	INF*	PTZ*	SUB1	NAM0
GELBER	ADJ1	ADV1	VRB*	INF*	PTZ*	SUB1	NAM0
GENERATIONENPROBLEM	ADJ0	ADV*	VRB*	INF*	PTZ*	SUB1	NAM0
GERAUCHT	ADJ*	ADV0	VRB0	INF*	PTZ1	SUB0	NAM0
GERUEHMTE	ADJ1	ADV0	VRB0	INP*	PTZ*	SUB0	NAM0

<S. 168>

G.I. (ZI) LERNWOERTERBUCM SEITE 2
29.04.1970

GESCHAEMT	ADJ*	ADV0	VRB0	INF*	PTZ1	SUB0	NAM0
GESCHOSS	ADJ*	ADV0	VRB*	INF*	PTZ*	SUB1	NAM0
GLOCKEN	ADJ0	ADV0	VRB0	INF0	PTZ*	SUB1	NAM0
GRINST	ADJ*	ADV-	VRB8	INF*	PTZ-	SUB-	NAM-
HEERFUEHRER	ADJ0	ADV0	VRB*	INF*	PTZ*	SUB1	NAM0
HELFT	ADJ*	ADV0	VRB1	INF*	PTZ0	SUB0	NAM0
HERZLICHE	ADJ1	ADV0	VRB0	INF*	PTZ*	SUB0	NAM0
KAFFEEPULVER	ADJ0	ADV0	VRB*	INF*	PtZ*	SUB1	NAM0
KIRCHTURM	ADJ*	ADV0	VRB*	INF*	PTZ*	SUB1	NAM0
KOHLENHAENDLER	ADJ0	ADV1	VRB*	INF*	PTZ*	SUB0	NAM0
KOSAKEN	ADJ1	ADV1	VRB1	INF0	PTZ*	SUB1	NAM0
KRAXELT	ADJ*	ADV-	VRB8	INF*	PTZ-	SUB-	NAM-
MACHTKOMPLEXES	ADJ1	ADV*	VRB*	INF*	PTZ*	SUB8	NAM-
MITTELGROSSEN	ADJ1	ADV0	VRB0	INF0	PTZ*	SUB0	NAM0
NEGERN	ADJ*	ADV0	VRB0	INF0	PTZ*	SUB1	NAM0
OSTERZEIT	ADJ*	ADV0	VRB0	INF*	PTZ-	SUB1	NAM0
PERUECKE	ADJ0	ADV0	VRB0	INF*	PTZ*	SUB1	NAM0
PEST	ADJ*	ADV-	VRB1	INF*	PTZ-	SUB8	NAM-
PHONEM	ADJ-	ADV*	VRB*	INF*	PTZ*	SUB8	NAM-
PROTESTIERTEN	ADJ0	ADV0	VRB1	INF0	PTZ*	SUB0	NAM0
TAPETE	ADJ0	ADV0	VRB0	INF*	PTZ*	SUB1	NAM0
UEBERHOEREN	ADJ0	ADV0	VRB0	INF1	PTZ*	SUB0	NAM0
VERBEUGE	ADJ-	ADV-	VRB8	INF*	PTZ*	SUB-	NAM-
VERKLEBT	ADJ*	ADV0	VRB1	INF*	PTZ0	SUB0	NAM0
WANGE	ADJ0	ADV0	VRB0	INF*	PTZ*	SUB1	NAM0
WETTERHAHN	ADJ*	ADV0	VRB0	INF0	PTZ*	SUB1	NAM0
WUNDERTEN	ADJ0	ADV1	VRB0	INF0	PTZ*	SUB0	NAM0
ZAEHFLUESSIGER	ADJ1	ADV0	VRB*	INF*	PTZ*	SUB*	NAM*
ZURUECKZUERHALTEN	ADJ0	ADV0	VRB1		INF0	PTZ*	SUB0 NAM0
EULENBAUM	ADJ*	ADV0	VRB*	INF*	PTZ*	SUB1	NAM0

<S. 169>

SCHNELLDRUCKERAUSGABE ALLER TESTSÄTZE

A) Konstruierte Testsätze mit vorgegebenen "unbekannten" Wortformen (Satznummern 20001 - 20283)

B) Zeitungsartikel-Sätze (Satznummern 500 - 556)

<S. 170>

TESTSAETZE SEITE 1

20001
GELBER , ZAEHFLUESSIGER HONIG VERKLEBT DAS GANZE PARADIES.

20002
GELBER DIE GLOCKEN NIE BLUEHEN ALS IN DER OSTERZEIT.

20003
BUNDESPOLITIKER WAREN SIE , SOLANGE SIE KEINE SOLDATEN BRAUCHTEN.

20004
IM VERGLEICH ZUM RASSISMUS IST DAS GENERATIONENPROBLEM KEIN ANALOGER , SONDERN DER EIGENTLICHE KONFLIKT.

20005
ES SAH AUS WIE DUNKELBRAUNES GROBES KAFFEEPULVER , UND ER HATTE NOCH NIE DAVON GERAUCHT.

20006
MAN SAGT , IN DER LETZTEN MINUTE HABE DER GEWALTIGE HEERFUEHRER SICH ANDERS BESONNEN UND SICH SEINES MACHTKOMPLEXES SEHR TIEF GESCHAEMT.

20007
WIEDER EINMAL HATTE DER ALLSEITS GERUEHMTE KUENSTLER EIN KLEINES STUECK EINER DUNKELGRUENEN TAPETE VON DER WAND SEINES SALONS GELOEST UND ZU EINEM ERSCHWINGLICHEN PREIS VERKAUFT.

20008
DIE GESELLSCHAFT ALS SOLCHE GEHOERT ZU DEN AM MEISTEN UND VON DEN MEISTEN BEANSTANDETEN UEBELN DER GESELLSCHAFT.

20009
NACH EINEM ERGREIFENDEN KONZERT IN EINER MITTELGROSSEN STADT BESTAND EINE BESUCHERIN DARAUF , DAS GELD FUER DEN EINTRITT ZURUECKZUERHALTEN , WEIL SIE ERFAHREN HATTE , DASS ALLE DEM CHOR ANGEHOERENDEN KOSAKEN AMERIKANER WAREN.

20010
ALS GEGEN ANFANG UNSERES JAHRHUNDERTS DIE EHRWUERDIGE SITTE VERBOTEN WURDE , NACH DER MAENNER UND FRAUEN , DIE IN SCHULD GEFALLEN WAREN , IHR HAUPT MIT ASCHE ZU BEDECKEN HATTEN , PROTESTIERTEN DIE FRISEURE UND KOHLENHAENDLER.

20011
IN LANGEN EROERTERUNGEN ENTSCHLOSS SICH DER PARTEIVORSTAND , DEN VORWURF ZU UEBERHOEREN , DIE GESINNUNG IHRES VORSITZENDEN SEI DUNKELBRAUN , UND HOB DAS HERZLICHE VERHAELTNIS ZU DEN NEGERN IN DER BEVOELKERUNG HERVOR.

20012
SEINEN EINDRUCK VCM SPIEGELSAAL BESCHRIEB DER STAATSMANN MIT DEN WORTEN , ER HABE DAS GEFUEHL GEHABT , DASS IHN VON ALLEN SEITEN BERUEHMTE STAATSMAENNER ANLAECHELN.

20013
GRINST NICHT SO VERLEGEN , SONDERN HELFT MIR DIE PERUECKE WIEDERZUFINDEN!

20014
MIT DIESER OHRFEIGE AUF DIE ANDERE WANGE HAT DER JUNGE PRIESTER EIN HISTORISCHES VERSPRECHEN EINGELOEST.

<S. 171>

TESTSAETZE SEITE 2

20015
NACHDEM ER AUSFUEHRLICH UEBER DIE VERHEERENDEN FOLGEN DER PEST GESPROCHEN HATTE , STIMMTEN ALLE ANWESENDEN DEM ANTRAG ZU , DER POLIZEI MEHR RECHTE GEGENUEBER DEMONSTRIERENDEN STUDENTEN
ZU GEBEN.

20016
BEI ALLEM TASTENDEN FORTSCHRITT INS UNGEWISSE STEHT DIE KATHOLISCHE KIRCHE FELSENFEST.

20017
VOR GERICHT SAGTE ER , DIE AXT HAETTE IHM BESSER GELEGEN ALS EINE PISTOLE.

20018
BEGAFFT UNS NICHT SO!

20019
WENN IHR AUSGERUHT HABT , KRAXELT IHR AUF DEN NAECHSTEN KIRCHTURM UND DREHT DEN WETTERHAHN NACH DEM WIND.

20020
SCHON SEIT ENDE DES ZWEITEN WELTKRIEGES WUENSCHTE SICH DIE JUNGE FRAU AUS BERLIN EINEN MANTEL AUS DUNKELBRAUNEM LEDER.

20021
WENN ES STIMMT , DASS DIE KLEINSTE , SINNUNTERSCHEIDENDE EINHEIT DER SPRACHE EIN PHONEM IST , SO GIBT ES LEUTE , DIE NOCH NIE EINES GESPROCHEN HABEN.

20022
ZU DEM GEWUENSCHTEN ERFOLG KAMEN SIE ERST AM ENDE DES DRITTEN ANLAUFS.

20023
ALS DIE DREI IM FAHRSTUHL SICH WUNDERTEN , DASS NACH DEM ZEHNTEN GESCHOSS NICHT , WIE ERWARTET , DAS ELFTE FOLGTE , GING ES SCHON BERGAB.

20024
EIN ABDRUCK KOMMT BILLIGER.

20025
ICH EMPFINDE ALLMAEhLICH EINE LEICHTE ABNEIGUNG GEGEN DUNKELBRAUNE TOENE.

20026
NICHT JEDER BEAMTE DER MITTLEREN LAUFBAHN IST NACH FUENF DIENSTJAHREN EIN ANGSTHASE.

20027
DAS PASSIERT MIR IMMER , WENN ICH MICH VERBEUGE.

20028
DORT SITZT EIN GELBER SCHMETTERLING.

20029
HATTE ER NICHT SCHON IMMER BUNDESPOLITIKER WERDEN WOLLEN*

20030
DAS HAETTE IN ANALOGER FORM GEMACHT WERDEN SOLLEN!

<S. 172>

20031
MEINE FRAU , DIE MANCHMAL AUF DUNKELBRAUNES LAND TRITT , LIEBT BLAUE NELKEN.

20032
ES GIBT LEUTE DIE UNTER DEN FOLGEN EINES MACHTKOMPLEXES LEIDEN.

20033
DUNKELGRUENEN DRECK FRASS WILLI IM URWALD.

20034
DIE BEANSTANDETEN PFLICHTUEBUNGEN AN DER UNI WURDEN ABGELEHNT.

20035

KOSAKEN , WELCH EIN WORT!

20036
WIR BEDECKEN UNSERE SCHWARZEN KLEINEN FUESSE.

20037
SCHON WIEDER DIE FARBE DUNKELBRAUN.

20038
OBSCHON WIR UNS OFT ANLAECHELN , ZUM LACHEN IST DIE SYNTAX NICHT!

20039
WER GRINST , TOETET NICHT.

20040

EINGELOEST WURDE SEIN VERSPRECHEN NIE!

20041
WIE DIE PEST WIRKTE DAS MITTEL.

20042
BILDEN SIE NOCH EINEN SATZ MIT ' FELSENFEST '

20043
IM WALD DIE AXT ERSETZT DIE FLINTE.

20044
BEGAFFT IHN NUR!

20045
UND DABEI KRAXELT ER NOCH GANZ GUT.

20046
OB GRUEN , OB GELB , IN DUNKELBRAUNEM SAMT GEHUELLT IST MAN EIN KOENIG.

20047
EIN PHONEM IST WICHTIG.

20048
ER WAR WEGEN SEINES EINZIGARTIGEN ANLAUFS BERUEHMT GEWORDEN.

20049
BERGAB GEHT ES SCHNELLER.

<S. 173>

TESTSAETZE SEITE 4

20050
DER ZAHNARZT - JAWOHL - MACHT AUCH MANCHMAL EINEN ABDRUCK.

20051
DUNKELBRAUNE ROSEN LIEBT MAN IN TIROL.

20052
ANGSTHASE NENNT MAN MISTER EULENBAUM.

20053
LIEBER ZIMMERMANN , VERBEUGE DICH VOR UNS!

20054
DIE EXPANSION GELBER CHINESEN NIMMT TAEGLICH ZU.

20055
BUNDESPOLITIKER SIND VIELLEICHT AUCH MENSCHEN.

20056
EIN ZU DIESEM PROBLEM ANALOGER SACHVERHALT WAERE NOCH ZU UEBERPRUEFEN.

20057
DIE DAME FUEHRT EIN DUNKELBRAUNES HUENDCHEN SPAZIEREN.

20058
ER MUSS ZUR SANIERUNG SEINES MACHTKOMPLEXES EINEN PSYCHOTHERAPEUTEN KONSULTIEREN.

20059
DIE BAEUME MIT DEN DUNKELGRUENEN BLAETTERN SIND HOHL.

20060
DIE BEANSTANDETEN LIEFERUNGEN WERDEN ZURUECKGEBRACHT.

20061
WER WEISS , WAS KOSAKEN FUEHLEN.

20062
BEDECKEN SIE IHR KNIE!

20063
DUNKELBRAUN SIND ALLE MEINE KLEIDER!

20064
DAS ANLAECHELN DES PERSONALS IST VERBOTEN.

20065
IHR GRINST NIE.

20066
ER HAT NACH MANCHEM AUF UND AB WAEHREND SEINES LEBENS KURZ VOR SEINEM TODE DEN SCHULDSCHEIN AUS SEINEN JUGENDTAGEN EINGELOEST.

20067
DIE PEST IST VERGESSEN.

20068
PETRUS STEHT FELSENFEST.

<S. 174>

20069
WER HAT DIE AXT GESTOHLEN*

20070
BEGAFFT IHR NICHT DIE TIERE IM ZOO.

20071
IHR KRAXELT AUF DEN HOEHEN DES BERGES , BIS IHR RUNTERFALLT.

20072
DUNKElBRAUNEM KAFFEE TRAUE NICHT.

20073
EIN PHONEM KOMMT SELTEN ALLEIN.

20074
ZWECKS ANLAUFS STUERZTE ER SICH IN DEN SEE.

20075
BERGAB FAELLT ES SICH SCHNELLER.

20076
EIN ABDRUCK SPIELT KEINE ROLLE.

20077
DUNKELBRAUNE AUGEN ZEUGEN VON EINEM SCHLECHTEN CHARAKTER.

20078
DER MANN ALS SOLCHER IST EO IPSO EIN ANGSTHASE.

20079
VERBEUGE DICH NIE.

20080
ER KONNTE SICH NICHT DARAN ERINNERN , OB DER VOGEL EIN GELBER WAR.

20081
WIR WISSEN NICHT , WAS DIE BUNDESPOLITIKER VORHABEN.

20082
DIES IST EIN ANALOGER FALL ZU DEM VORHERIGEN.

20083
EIN DUNKELBRAUNES ETWAS LAEUFT UEBER DEN WEG IN DEM MOMENT , IN DEM ICH ZU SINGEN ANFANGE.

20084

ER HAT IHN , ALS HAETTE ER NIE SO ETWAS SCHOENES GESEHEN, BEGAFFT DEN GANZEN TAG , SOGAR MIT WACHSENDER BEGEISTERUNG.

20085
DIE FAKTOREN DES MACHTKOMPLEXES SIND DURCHAUS ANALYSIERBAR.

20086
SIE KAM DOCH VOM SOMMERSCHLUSSVERKAUF MITDUNKELGRUENEN SCHUHEN ZUURUECK, OBWOHL SIE SCHON ZUR GENUEGE MIT DUNKELGRUENEN SCHUHEN EINGEDECKT IST.

<S. 175>

20087
VERSCHMITZT GRINST FRITZ ALS DEM FENSTER HERAUS UND TUT SONST SO , ALS WUESSTE ER VON NICHTS.

20088
NACH MONATELANGEN REPARATURARBEITEN WAREN DIE BEANSTANDETEN MAENGEL IMMER NCCH NICHT BEHOBEN.

20089
DIE LIEDER DER KOSAKEN GEHEN MIR STAENDIG IM KOPF HERUM.

20090
SCHNEEFELDER BEDECKEN DIE GIPFEL DES HOCHGEBIRGES , BLUMENFELDER DIE TAELER.

20091
MIT DUNKELBRAUN GEFAERGTEN HAAREN , NEUEN ZAEHNEN UND EINEM MASSGESCHNEIDERTEN ANZUG ERSCHIEN DER LAENGST VERGESSENE PLOETZLICH UND UNERWARTET ZUM FRUEHSTUECK BEI TIFFANY.

20092
ANLAECHELN KOSTET NICHTS.

20093
SIE HATTE DEN GUTSCHEIN , DEN ZU ERLANGEN SIE GROSSE MUEHE AUFGEWENDET HATTE , NICHT RECHTZEITIG - UND DAMIT PRAKTISCH UEBERHAUPT NICHT - EINGELOEST , SO DASS SIE DIE REISE NICHT BEKAM.

20094
ER GLAUBT SO FELSENFEST , SO UEBERZEUGT AN EINEN GEWINN , OBWOHL ER NICHT MITGESPIELT HAT.

20095
DA DIE AEXTE BILLIGER GEWORDEN SIND WOLLEN AUCH WIR UNS HEUTE EINE AXT KAUFEN.

20096
ICH KANN NICHT VERSTEHEN , WARUM SO EINER SOGAR WOCHENLANG KRAXELT , UM AUF EINE BERGSPITZE ZU GELANGEN UND DANN DOCH MUEDE UND - ES MUSS EINMAL GESAGT WERDEN - MIT ZWEIFELHAFTEM ERFOLG WIEDER MIT DEM ABSTIEG ZU BEGINNEN.

20097
DIE FLASCHE MUSSTE MIT DUNKELBRAUNEM SAFT GEFUELLT WERDEN , DAMIT DER BETRUG MOEGLICHST LANGE VERHEIMLICHT WUERDE.

20098
DAS EINZELNE PHONEM KANN NUR IM HINBLICK AUF DAS GESAMTE PHONEMSYSTEM BESCHRIEBEN WERDEN.

20099
DIE ZUSCHAUER HATTEN NICHT MEHR MIT DEM GROSSEN ERFOLG DES DRITTEN ANLAUFS GERECHNET.

20100
MIT SEINER GESUNDHEIT GING ES SEHR SCHNELL BERGAB.

20101
IM VERZEICHNIS AUF DER POLIZEISTELLE HATTEN SIE SCHNELL SEINEN ABDRUCK GEFUNDEN.

<S. 176>

20102
ICH TRAUTE MEINEN AUGEN NICHT , ALS BEIM BETRETEN DES WALDE S DER DUNKELBRAUNE ANGSTHASE AUF MICH ZUKAM.

20103
ER HATTE IHN ANGSTHASE GENANNT , UM SEINEN MUT ANZUSTACHELN.

20104
MAN SAGT VON IHM , ER VERBEUGE SICH VOR ALLEN , VON DENEN ER MEINT , DASS SIE IHM NUETZLICH SEIN KOENNTEN.

20105
IM MITTELALTER STARBEN VIELE MENSCHEN AN DER PEST.

20106
WARUM WIRD DAS SONNENLICHT IM LAUFE DER JAHRMILLIONEN NICHT GELBER*

20107
BRANDT IST BUNDESPOLITIKER.

20108

IN AMERIKA UND EUROPA BEDIENT MAN SICH GANZ ANALOGER COMPUTER.

20109
WER TRINKT NICHT LIEBER DUNKELBRAUNES ALS HELLES BIER*

20110
WENN DIE AUSMASSE DIESES MACHTKOMPLEXES SICHTBAR WERDEN , WIRD DER WELTFRIEDE GEFAEHRDET SEIN.

20111
MAN BEHAUPTET OFT , DER MARS WAERE VON KLEINEN DUNKELGRUENEN
MAENNCHEN BEWOHNT.

20112
DIESE LEUTE BEANSTANDETEN ABER SEHR DAS FEHLEN VON SAUERSTOFF , WENN SIE DA LEBTEN.

20113
JEDENFALLS IST GANZ SICHER , DASS ES KEINE KOSAKEN AUF DEM MARS GIBT.

20114
DA ES NUR WENIG WASSER GIBT , BEDECKEN WOHL NUR KLEINE FLECHTEN DEN KARGEN BODEN.

20115
DUNKELERAUN LEUCHTEN DIE VEGETATIONSARMEN GEBIETE IM SOMMER , WAEHREND SIE IM WINTER EINE HELLERE TOENUNG ZEIGEN.

20116
WIE MAG ES IN DIESER EINSAMKEIT AUSSCHAUEN , WENN EINEN DIE BEIDEN MONDE PHOBOS UND DEIMOS ANLAECHELN*

20117
SICHER GRINST PHOAUS FRECH , WEIL ER NOCH SO KLEIN IST UND SICH NICHT BENEHMEN KANN.

<S. 177>

20118
HAST DU DEIN VERSPRECHEN EINGELOEST , UNS ETWAS UEBER DIE KRANKHEITSERREGER AUF DEM MARS ZU BERICHTEN*

20119
JA , DIE FLECHTEN WERDEN ERSTAUNLICH OFT VON EINER ART VON PEST BEFALLEN.

20120
SIE WACHSEN DANN FELSENFEST ZUSAMMEN.

20121
DAS IST EIN GEWAECHS , DAS NUR MIT AXT UND BEIL ZERSCHLAGEN WERDEN KANN.

20122
DA STAUNT IHR WOHL , DA BEGAFFT IHR MICH.

20123
KRAXELT NICHT AUF DEM SOFA HERUM , WENN ICH EUCH ETWAS ERZAEHLE.

20124
MAN SIEHT ZWAR DUNKELBRAUNEM STOFF NICHT AN , WIE SEHR IHR IHN STRAPAZIERT , ABER HOERT JETZT ENDLICH AUF.

20125
WER HAT DIESES UNANSTAENDIGE PHONEM VON SICH GEGEBEN*

20126
ES BEDARF NUR EINES KLEINEN ANLAUFS , BIS ICH RICHTIG IN FAHRT KOMME UND EUCH VERHAUE

20127
WIR GINGEN LANGSAM BERGAB , ALS WIR AUF EINE INTERESSANTE SPUR STIESSEN.

20128
ES HANDELTE SICH ZWEIFELLOS UM DEN ABDRUCK EINER PFOTE.

20129
DER DUNKELBRAUNE BODEN LIESS UNS DIE FAEHRTE NUR SCHWER ERKENNEN.

20130
SCHLIESSLICH WAREN WIR UNS EINIG , DASS NUR EIN TIER , NAEMLICH EIN ANGSTHASE , SOLCHE SPUREN HINTERLASSEN KANN.

20131
ICH VERBEUGE MICH VOR DEM ALLWISSENDEN COMPUTER.

20132
AM ABEND SCHWINGEN DIE KOSAKEN SAEBEL UND TANZEN.

20133
DER ABDRUCK SEINER SCHUHSOHLEN HAT IHN VERRATEN.

20134
DIE PRUEFER BEANSTANDETEN DEN ZUSTAND DER VERSCHNUERLNG UND DIE UNUEBLICHE ART UND WEISE DER HALTERUNGEN.

20135
DIE BEANSTANDETEN PAPIERE HAST ALSO DU EINGELOEST!

20136
BERGAB ROLLEN UND RUTSCHEN STEINE UND GEROELL.

20137
PHONOLOGISCHE MERKMALE ORDNEN EIN PHONEM EINER BESTIMMTEN
KLASSE ZU , WAEHREND PHONETISCHE MERKMALE PHONETISCHE SKALEN BEZEICHNEN , DEM PHONEM KOMMT ALSO HIER GROESSERE BEDEUTUNG ZU.

20138
ICH ANGSTHASE!

20139
DU BIST DAVON FELSENFEST UEBERZEUGT , DASS ER GEGEN ABEND KOMMEN WIRD.

20140
ANGESICHTS DIESES MACHTKOMPLEXES , DER SICH STAENDIG VERGROESSERT UND VERHINDERT , DASS EINE BESCHAEFTIGUNG MIT DEN TIEFERLIEGENDEN URSACHEN ENDLICH ERMOEGLICHT WIRD , IST GEMEINSAMER WIDERSTAND DAS GEBOT DER STUNDE.

20141
DIE MENGE HAT IHR SCHAUSPIEL UND BEGAFFT NOCH IMMER , OBWOHL LAENGST NICHTS MEHR ZU SEHEN IST , DIE STRASSENKREUZUNG.

20142
WEISSE SCHNEEFLOCKEN FALLEN VOM HIMMEL UND BEDECKEN WALD UND FLUR.

20143
ER KONTROLLIERT NOCHMALS SEINE WERKZEUGE , SAEGE UND AXT.

20144
TROTZ DES GROSSEN ANLAUFS PURZELTE ER ZU BODEN.

20145
EIN ANALOGER SACHVERHALT ERGIBT GLEICHE STRAFEN.

20146
UND ICH BESCHLOSS , BUNDESPOLITIKER ZU WERDEN.

20147
IM MITTELALTER DROHTEN PEST UND HUNGER GANZE LANDSTRICHE ZU ENTVOELKERN.

20148
DER SENNWIRT KRAXELT IN SEINER FREIZEIT.

20149
SIE KOENNEN MICH NOCH SO CHARMANT ANLAECHELN , IHRE ZEUGNISSE BLEIBEN DIE GLEICHEN.

20150
OBWOHL ICH MICH EHRERBIETIG NAEHERE UND , NACH LANDESSITTE , EINMAL PRO TREPPENSTUFE VERBEUGE , FAELLT KEIN WOHLWOLLENDER BLICK AUF MICH , DER MIR DIE PROZEDUR EINIGERMASSEN ERLEICHTERT HAETTE.

<S. 179> TESTSAETZE SEITE 10

20151
DIE WESTE HAT EINEN STICH INS DUNKELBRAUNE.

20152
' JE GELBER DESTO LIEBER ' IST SEINE DEVISE.

20153
SEINE PUPILLEN SCHIMMERN IN EINEM FORT DUNKELBRAUN.

20154
ER FOLGT MANCH DUNKELBRAUNEM VERDAECHTIGEN.

20155
GRINST LIEBER NICHT , WENN ER KOMMT.

20156

20157
ICH WILL KEIN DUNKELGRUENES KLEID , LIEBER WAERE MIR EIN DUNKELBRAUNES.

20158
DIESE DUNKELGRUENEN VORHAENGE SIND FUER DIE AUGEN DAS ALLERBESTE.

20159
ICH HABE NIE EIN GESICHT GELBER WERDEN SEHEN.

20160
WAS WAERE KOMISCHER ALS BUNDESPOLITIKER AUF DEM REDNERPULT*

20161
WIR KOENNEN ALSO FESTSTELLEN , DASS EIN DAZU WEITGEHEND ANALOGER FALL IM ALKOHOLBEDINGTEN LEBERSCHWUND VORLIEGT.

20162
EIN DUNKELBRAUNES TRITT MIT LEISEN SCHRITTEN INS ZIMMER.

20163

HITLERS FRAUENBEKANNTSCHAFTEN AUF DEM GIPFEL SEINES MACHTKOMPLEXES - EIN HOECHST ABSTOSSENDES THEMA.

20164
ZWISCHEN EINEM GRAUEN ZEBRA UND EINEM HELLGRUENEN BESTEHT EIN KLEINERER UNTERSCHIED ALS ZWISCHEN EINEM BRAUNEN UND EINEM
DUNKELGRUENEN.

20165
IN DER ZWAR NICHT VON DER GESAMTEN KOMMISSION , WOHL ABER EINIGEN IHRER MITGLIEDER BEANSTANDETEN PASSAGE WURDE DAS WORT ' ASCH ' GESTRICHEN.

20166
WARUM KOSAKEN , EBENSOWENIG RUSSEN WIE DIE TATAREN , ZU DEN ERBITTERTSTEN FEINDEN DES STALINISMUS ZAEHLTEN , IST JEDEM KENNER DER VERHAELTNISSE KLAR.

20167
BEDECKEN SIE EINMAL IHRE BLOESSE MIT EINER UNTERHOSE!

<S. 180> TESTSAETZE SEITE 11

20168
OB DUNKELBRAUN AUF AFFEN ANZIEHENDER WIRKT ALS BANANEN AUF
EINEN ESEL , KANN BEIM STAND DER FORSCHUNG NOCH NICHT GESAGT WERDEN.

20169
SICH AUF DER STRASSE , IM CAFE , IM PUFF , IM BUERO VON EINEM BLONDEN , EINEM ROTEN ODER AUCH EINEM UNBEKLEIDETEN MAEDCHEN ANLAECHELN ZU LASSEN , IST ALLEIN NICHT BEFRIEDIGEND,

20170
WEIL , WER , WENN ER EINEN , DER IHM UEBERGEORDNET IST , IN UNTERHOSEN SIEHT , GRINST ODER BLOED LACHT , SPINNT , IST IHM ZU VERZEIHEN.

20171
EINGELOEST BEDEUTET DER SCHECK NICHTS ALS EIN STUECK GRUENES PAPIER.

20172
ERWIN HAT , WAS MAN NIE GEGLAUBT HAETTE , PEST UND CHOLERA ZUGLEICH.

20173
MAX VERTRITT SEINE MEINUNG , FELSENFEST UND UNERSCHUETTERLICH ALLEN VERNUENFTIGEN ARGUMENTEN UNZUGAENGLICH.

20174
MIT WUCHTIGEM SCHRITT , DIE LIPPEN FEST VERBISSEN , UNGEWASCHEN , DEN HUT TIEF IN DER STIRN , AXT UND SAEGE UEBER DER SCHULTER , STAMPFT DER HOLZFAELLER , FINSTRE PLAENE SCHMIEDEND , HEIM ZU SEINER FRAU.

20175
DER NEUE WIRD , WIE IN DIESEM VIERTEL NICHT ANDERS UEBLICH , ZUNAECHST EINMAL , ALS SEI ER VOM MOND GEFALLEN , AUSGIEBIG BEGAFFT.

20176
KARL KRAXELT UND KRAXELT , UND ER KOMMT SCHLIESSLICH OBEN AN.

20177
MIT SCHWARZEM , DUNKELBRAUNEM , MIT BLONDEM , ROTEM , BRUENETTEM , MIT UND OHNE HAAR HAB ICH SIE GELIEBT.

20178
PHONEM UND MORPMEM NENNEN DIE STRUKTURALISTEN ALS GRUNDEINHEITEN DER SPRACHE.

20179
WEGEN DEINES UNORDENTLICHEN ANLAUFS KOMMST DU NIE UEBER DIE MAUER.

20180
VON NUN AN GING ES , WIE DIE KNEF , DIESE ALTE VETTEL , SO SCHOEN SINGT , BERGAB.

20181
DER , DER DEN KEINEN ABDRUCK HINTERLASSENDEN HANDSCHUH TRUG , WAR ES.

<S. 181>

20182
DIE LINKE OBERE ECKE SPIELT VON EINEM LEICHTEN BLAU INS DUNKELBRAUNE.

20183
ANGSTHASE NENNEN HANS HOECHSTENS DIE , DIE IHN DAMALS NICHT AUF DEN WALDAIHOEHEN GFSEHEN HABEN.

20184
WER NOCH EINEN FUNKEN AN EHRGEFUEHL IM LEIB HAT , VERBEUGE SICH VOR DIESEM MANN.

20185

MIT DER VERSTAERKTEN EINFUHR GELBER ROSEN VERSUCHT MAN JETZT , DER WACHSENDEN NACHFRAGE GERECHT ZU WERDEN.

20186
DIESER MANN HAT ALS BUNDESPOLITIKER EINE RICHTUNG EINGESCHLAGEN , DIE , SOBALD SIE VOM WAEHLER RICHTIG EINGESCHAETZT WIRD , DER PARTEI STIMMEN KOSTEN KANN.

20187
DEM RASCHEN ERKENNEN ANALOGER BEZIEHUNGEN SIND WIR MIT DIESEN VERFAHREN EIN GUTES STUECK NAEHER GEKOMMEN.

20188
DIE SCHAUSPIELERIN TRUG EIN SCHLICHTES DUNKELBRAUNES KLEID.

20189
DAS VERHALTEN DES STAATSMANNES DUERFTE DIE FOLGE EINES MACHTKOMPLEXES SEIN.

20190
DER JAEGER KANN SEINEN ANZUG , DER DUNKELGRUENEN FARBE WEGEN , NICHT ZUR BEERDIGUNG ANZIEHEN.

20191
BEANSTANDETEN WAREN GEHOERT IN UNSEREM GESCHAEFT KEIN PLATZ MEHR.

20192
DER KAMPF ZWISCHEN KOSAKEN UND TATAREN ABER SEI SO ERBITTERT GEFUEHRT WORDEN , DASS NUR FEIGLINGE UNVERLETZT BLIEBEN.

20193
BEDECKEN WIR IHN MIT DEM FEDERBETT , UND LASSEN WIR IHN WEITERSCHLAFEN!

20194
DER BLASSE URLAUBER BEKAM AM ZIELORT ZWAR EINEN SONNENBRAND , KEHRTE ABER DENNOCH DUNKELBRAUN ZURUECK.

20195
ANLAECHELN ALLEIN IST UNBEDEUTEND.

20196
DER DORT GRINST IMMER SO , WENN ER GLAUBT , SEINEN PARTNER UEBERVORTEILT ZU HABEN.

20197
WESHALB HAST DU DEN SCHECK EINGELOEST , OBWOHL DER AUSSTELLER WEGEN SEINER UNGEWOEHNLICHEN GESCHAEFTSGBAREN KEINEN GUTEN RUF GENIESST*

<S. 182>

20198
DER ' SCHWARZE TOD ' , DIE PEST , KANN IN LAENDERN MIT GEWISSEN HYGIENISCHEN EINRICHTUNGEN ALS BESIEGT BETRACHTET WERDEN.

20199
DIE INTERESSENGEMEINSCHAFT GLAUBT FELSENFEST AN EINE BESSERE
ZUKUNFT IN FINANZIELLER HINSICHT.

20200
AXT UND BEIL SIND NICHT DIE GLEICHEN WERKZEUGE.

20201
WARUM WIRD DIESER ARME MENSCH IMMER WIEDER SO AUFFAELLIG BEGAFFT*

20202
NOCH IN SEINEM HOSEN ALTER KRAXELT DER VON VIELEN FERNSEHZUSCHAUERN GELIEBTE BERGSTEIGER UND ERZAEHLER IN DEN ALPEN MERUM.

20203
DUNKELBRAUNM STOFF WIRD IN DER NAECHSTEN SAISON KAUM INTERESSE BESCHIEDEN SEIN, WENN MAN DER PROGNOSE PROMINENTER MODESCHOEPFER GLAUBEN SCHENKEN DARF.

20204
EIN LAUT , WELCHER IN DERSELBEN STELLUNG EINEN BEDEUTUNGSUNTERSCHIED HERVORRUFT , HEISST PHONEM.

20205
DIE KRAFT DES ANLAUFS IST BEIM RENNRODELN NICHT WENIGER WICHTIG ALS DAS ANHEIZEN DER KUFEN.

20206
EIN VOLKSWAGEN KANN , SOBALD ES BERGAB GEHT , UNGEAHNTE GESCHWINDIGKEITEN ERREICHEN.

20207
DIE VERBRAUCHERZEITSCHRIFT GESTATTET DEN ABDRUCK IHRER ARTIKEL UNENTGELTLICH , WENN DIESER OHNE JEDE KUERZUNG ODER HINZUFUEGUNG ERFOLGT UND UNTER ANGABE DER QUELLE VEROEFFENTLICHT WIRD.

20208
DUNKELBRAUNE AUGEN FINDET MAN BEI MENSCHEN MIT HELLBLONDEN
HAAREN NUR SEHR SELTEN.

20209

ES WAR IHM SEHR PEINLICH , WENIGE TAGE VOR DER VERLEIHUNG EINER AUSZEICHNUNG ALS ANGSTHASE HINGESTELLT ZU WERDEN.

20210
ICH VERBEUGE MICH VOR DEM , DER GLAUBT , ER KOENNE MIT DIESEN SAETZEN TATSAECHLICH ETWAS ANFANGEN.

20211
ER UEBERREICHTE DER FRAU DES PRAESIDENTEN EINEN STRAUSS GELBER ROSEN.

20212
BUNDESPOLITIKER ALLER LAENDER VEREINIGT EUCH!

20213
DIE BILDUNG ANALOGER DIFFERENTIALGLEICHUNGEN , NACH AEHNLICHEN
GESETZEN AUFGESTELLT , DUERFTE KEINE SCHWIERIGKEITEN BEREITEN.

<S. 183> TESTSAETZE SEITE 14

20214
EIN DUNKELbRAUNES TUCH WAR FUER IHN DAS LETZTE , HOFFNUNGSVOLLE ZEICHEN , DAS ER VON SEINEM KIND , DAS INS AUSLAND ENTFUEHRT WORDEN WAR , ERHALTEN HATTE.

20215
DIE AUSDEHNUNG DES MACHTKOMPLEXES VERLIEF OHNE WEITERE HINDERNISSE.

20216
DIE DAME MIT DEM DUNKELGRUENEN HUT , EIN BERUEHMTES GEMAELDE , WURDE BEI DEM TRANSPORT ZU EINER GEMAELDEAUSSTELLUNG IM LOUVRE IN PARIS LEICHT , ABER DENNOCH SICHTBAR BESCHAEDIGT.

20217
ES GAB KAUM EINEN VORSCHLAG , DEN DIE OPPOSITIONSPARTEIEN NICHT BEANSTANDETEN.

20218
IM KOSTUEM EINES KOSAKEN BETRAT EIN UNBEKANNTER DIE BUEHNE , UM IWAN REBROW ZU IMITIEREN , WAS IHM ABER NICHT GELANG , SO DASS ER UNTER DEM ZISCHEN UND PFEIFEN DES PUBLIKUMS DIE BUEHNE VERLASSEN MUSSTE.

20219
VOM STURM GEPEITSCHTE WOLKENFETZEN BEDECKEN DEN HIMMEL , GRAU IN GRAU , UNHEILVERHEISSEND ZIEHT EIN GEWITTER AUF.

20220

DUNKELBRAUN - DIE MODEFARBE DES WINTERS NEUNZEHNHUNDERTSIEBZIG.

20221
WENN SIE KINDERAUGEN ANLAECHELN , IST SIE GLUECKLICH.

20222
ES GRINST DER TOD UEBER DIE VERHEERENDEN AUSMASSE DER PEST , HERVORGERUFEN DURCH DIE FOLGEN EINES LANGJAEHRIGEN KRIEGES.

20223
HABT IHR EUREN SCHECK EINGELOEST*

20224
FELSENFEST HEISST NICHT UNZERSTOERBAR.

20225
DIE AXT IM HAUSE ERSETZT DEN ZIMMERMANN.

20226
EIN AFFE BEGAFFT DIE MENSCHEN , ODER IST ES UMGEKEHRT*

20227
EINE AFFENHERDE KRAXELT MIT LAUTEM GESCHREI AUF DAS DACH DES HOTELS , SO DASS MAN GLAUBEN KCENNTE , EIN AUFRUHR SEI IN DER SONST SO RUHIGEN STADT AUSGEBROCHEN.

20228
IN DUNKELBRAUNEM ANZUG WURDE ER ZUERST GESEHEN.

20229
EIN PHONEM IST DIE KLEINSTE EINHEIT DER PHONOLOGIE.

20230
DIE TECHNIK DES ANLAUFS IST FUER VERSCHIEDENE SPORTLICHE DISZIPLINEN NICHT ZU UNTERSCHAETZEN.

20231
DER ABDRUCK SEINER ABHANDLUNG IN EINER WISSENSCHAFTLICHEN ZEITSCHRIFT , WAS REIN ZUFAELLIG WAR , OEFFNETE IHM ALLE TUEREN.

20232
DER AUFSTIEG VERLIEF OHNE SCHWIERIGKEIT , ABER BERGAB ERLEBTEN WIR EINE UEBERRASCHUNG.

20233
DUNKELBRAUNE GESTALTEN NAEHERTEN SICH IM DUNKELN.

20234
ANGSTHASE , UEBERWINDE DEINE FURCHT!

20235
VERBEUGE DICH , WENN DU ALS ZEUGE AUFGERUFEN WIRST!

20236
BEI DIESER BITTEREN KAELTE SPRANG DER VOLKSWAGEN , TROTZ DES LANGEN ANLAUFS , BERGAB NICHT AN.

20237
DU HAST DAS PHONEM ABER FALSCH AUSGESPROCHEN.

20238
ER BENIMMT SICH WIE DIE AXT IM WALDE.

20239
DER MANTEL AUS DUNKELBRAUNEM LEDER WAR EIN STUECK VON GANZ BESONDERER ELEGANZ UND SCHOENHEIT.

20240
ICH VERBEUGE MICH IN EHRFURCHT VOR MEINER MIUTTER.

20241
ES GIBT SITUATIONEN , BESONDERS IM KRIEG , IN DENEN SICH KEINER ZU SCHAEMEN BRAUChT , EIN ANGSTNASE ZU SEIN.

20242
DER SCHLECHT GERATENE SOHN GRINST SEINEM VATER INS GESICHT.

20243
TROTZ SEINES MACHTKOMPLEXES WAR ER EIN BEIM VOLK BELIEBTER STAATSMANN.

20244
MAN ERKANNTE SOFORT AN IHRER TRACHT , DASS ES KOSAKEN GEWESEN SIND.

20245
WENN KLEINE MAEDCHEN IHREN VATER ANLAECHELN , WIRD IHNEN SOGLEICH JEDER WUNSCH ERFUELLT.

20246
GELBER , DICKER QUALM BREITETE SICH UEBER DER HUETTENSTADT AUS.

20247
DER ERFAHRENE BUNDESPOLITIKER UNTERHIELT SICH MIT DER JUGEND
UEBER FRAGEN DER BILDUNGSPOLITIK.

20248
EIN DUNKELBRAUNES MAEDCHEN MIT WUNDERSCHOENEN GROSSEN

AUGEN SPIELTE MIT UNSEREM KLEINEN SOHN IM GARTEN.

20249
EIN ANALOGER VORFALL WAR IN DER GESTRIGEN AUSGABE DER FRANKFURTER ALLGEMEINEN ZU LESEN.

20250
ZU DEM DUNKELGRUENEN KOSTUEM WUERDE ICH DIR EMPFEHLEN , DIESEN HERRLICH GEMUSTERTEN SCHAL ZU TRAGEN.

20251
VIELE KAEUFER BEANSTANDETEN MANGELNDE BEDIENUNG WAEHREND DES
WINTERSCMLUSSVEPKAUFS IN DEN WARENHAEUSERN.

20252
DUENNE SCHNEESCHICHTEN UND E1S BEDECKEN DIE KNOSPEN DER TULPEN UND NARZISSEN.

20253
DUNKELBRAUN IST DIE LIEBLINGSFARBE MEINER MUTTER.

20254
BEVOR SIE DIE REISE NACH SUEDAFRIKA ANTRATEN , MUSSTEN ALLE TEILNEHMER AN DER SAFARI SICH GEGEN PEST IMPFEN LASSEN.

20255
ER BLIEB STEIF UND FELSENFEST BEI SEINEM ENTSCHLUSS.

20256
SAEMTLICHE GUTSCHEINE WURDEN EINGELOEST.

20257
UNSERE KLEINE TOCHTER KRAXELT SCHON HOHE BERGE HINAUF.

20258
DIE ALTE DAME BEGAFFT DAS JUNGE MAEDCHEN , WELCHES MIT EINEM
MINIROCK DURCH DAS CAFE STOLZIERT.

20259
DER ABDRUCK EINES FUSSES WAR FUER DIE KRIPO EIN WICHTIGER HINWEIS.

20260
VIELE LEUTE GEHEN LIEBER BERGAUF ALS BERGAB.

20261
DIE DUNKELBRAUNE SCHUHCREME ROCH ENTSETZLICH STARK NACH TERPENTIN.

20262
DAS WAR EIN GELBER PORSCHE.

20263
OHNE BUNDESPOLITIKER GAEBE ES KEINE BUNDESPOLITIK.

<S. 186> TESTSAETZE SEITE 17

20264
DAS FORMULAR IST IN ANALOGER WEISE ZU VERVOLLSTAENDIGEN.

20265
WIR SAHEN ETWAS DUNKELBRAUNES , UNDEFINIERBARES SICH LANGSAM AUF UNS ZU BEWEGEN.

20266
DAS ZENTRUM DES MACHTKOMPLEXES WIRD VON EINER GRUPPE BEAMTER BEHERRSCHT.

20267
DER ZEUGE ERINNERTE SICH NUR AN EINEN DUNKELGRUENEN MANTEL.

20268
DIE BEANSTANDETEN DAS NICHT IM GERINGSTEN.

20269
KOSAKEN REITEN NUR IM CIRCUS.

20270
DANN GLAUBTEN WIR NICHT MEHR DEN WAGEN MIT DER ZERRISSENEN PLANE BEDECKEN ZU KOENNEN.

20271
SPAETER GING DIE FARBE IN EIN TIEFES DUNKELBRAUN UEBER.

20272
ANLAECHELN ODER NICHT ANLAECHELN , DAS IST HIER DIE FRAGE.

20273
GRINST DU IMMER NOCH*

20274
NIE HABEN WIR UNSERE VERSPRECHUNGEN EINGELOEST.

20275
PEST UND BUDA LIEGEN EINANDER GEGENUEBER.

20276
NACH DEM GARTENFEST NUN AUCH NOCH EIN FELSENFEST!

20277
DIE AXT ER5PART DEN ZIMMERMANN.

20278

BEWUNDERT UND BEGAFFT : DAS NEUE MODELL UEBERTRAF ALLE ERWARTUNGEN.

20279
WAS KRAXELT DENN DA WOHL ZWISCHEN DEN KIESELN HERUM*

20280
DUNKELERAUNEM STAUB WAR NUR ETWAS ZUCKER BEIGEMENGT WORDEN.

20281
SO EIN PHONEM IST SCHON GANZ VORNEHM.

<S. 187> TESTSAETZE SEITE 18

20282
TROTZ LANGEM ANLAUFS GING ES IMMER NUR WEITER BERGAB , OHNE DASS DIESER ANGSTHASE DEN GERINGSTEN ABDRUCK HINTERLASSEN HAETTE.

20283
DUNKELBRAUNE SCHUHE , ER VERBEUGE SICH VOR DIESER GESCHMACKLOSIGKEIT!

<S. 188> TEST: ZEITUNGSARTIKEL SEITE 1

500
DIE BEI DER LETZTEN TARIFERHOEHUNG DER BUNDESBAHN AUSGEBLIEBENE ANHEBUNG DER AUFNAHMETARIFE FUER DIE SAAR / ZUM SCHUTZ DER ANSAESSIGEN STAHLINDUSTRIE / MUSS BIS ZUM ERSTEN SEPTEMBER 1971 NACHGEHOLT WERDEN.

501
DIES SIEHT EIN BESCHLUSS DER EUROPAEISCHEN KOMMISSION VOR , DER AM MONTAG DER BUNDESREGIERUNG UEBERMITTELT WURDE.

502
WIE ES IN DER BEGRUENDUNG HEISST , SEI MAN DAVON AUSGEGANGEN , DASS DIE AUSNAHMETARIFE NUR EINEN VORUEBERGEHENDEN CHARAKTER HABEN , UM DIE AUFRECHTERHALTUNG DER WEITBEWERBSTELLUNG DER SAARLAENDISCHEN STAHLINDUSTRIE BIS ZUM BAU EINES ANSCHLUSSES DER SAAR AN DIE SCHIFFAHRTSWEGE ZU SICHERN.

503
DURCH DIE NICHTBERUECKSICHTIGUNG DER AUSNAHMETARIFE DER LETZTEN TARIFERHOEHUNG BEI DER BUNDESBAHN HABE SICH DIE PRAEFERENZSPANNE JEDOCH ERHOEHT.

504
NACH EINER ENTSCHEIDUNG DER KOMMISSION VOM LETZTEN WOCHENENDE KOENNEN DIE AUSNAHMETARIFE FUER DIE

SAARWIRTSCHAFT BIS ZUM EINUNDDREISSIGSTEN DEZEMBER 1975 BEIBEHALTEN WERDEN.

505
IN EINER DARAUFFOLGENDEN VIERJAHRESPERIODE IST DER ABBAU DIESER SOGENANNTEN ALS-OB-TARIFE VORGESEHEN.

506
DIE USA WERDEN AUCH WEITERHIN EIN LAND OHNE PERSONALAUSWEISE BLEIBEN.

507
DAS JUSTIZMINISTERIUM HAT DEN PLAN FALLENGELASSEN , AUF BESONDEREN WUNSCH PERSONALAUSWEISE FUER PERSONEN AUSZUSTELLEN , DIE IN DEN GRENZGEBIETEN IN DER NAEHE VON KANADA ODER MEXIKO HAEUFIG DIE GRENZEN UEBERQUEREN MUESSEN.

508
BUERGERRECHTSORGANISATIONEN HATTEN GEGEN DIESEN PLAN PROTESTIERT , WEIL SIE DARIN DEN ERSTEN SCHRITT AUF DEM WEGE ZU EINEM ALLGEMEINEN AUSWEISZWANG IN DEN USA SAHEN.

509
IM ALLGEMEINEN WEIST MAN SICH IN DEN USA EINFACH MIT DEM FUEHRERSCHEIN AUS.

510
ES BEGANN VOR DREI JAHREN , ALS DIE FRANZOESISCHE REGIERUNG BONN MITTEILTE , DASS SIE NICHT MEHR IN DER LAGE SEI , GENUEGEND LEHRKRAEFTE ZUM FREMDSPRACHENUNTERRICHT INS FRANKOPHONE AFRIKA ZU SCHICKEN.

511
DIE BUNDESREGIERUNG WURDE GEBETEN , WENN SIE INTERESSE AN DER VERBREITUNG UND AM UNTERRICHT DER DEUTSCHEN SPRACHE HAETTE , AN DIE REGIERUNGSSCHULEN DER AFRIKANISCHEN LAENDER , IN DENEN FRANZOESISCH AMTSSPRACHE IST , EIGENE LEHRER ZU ENTSENDEN.

<S. 189> TEST: ZEITUNGSARTIKEL SEITE 2

512
DER DEUTSCH-FRANZOESISCHE FREUNDSCHAFTSVERTRAG UND DIE AKTIVIERUNG DER AUSWAERTIGEN KULTURPOLITIK MACHTEN EIN ENGAGEMENT DER DEUTSCHEN NOTWENDIG.

513
IM FRANKOPHONEN AFRIKA GIBT ES HEUTE UEBER FUENFZIG LEHRKRAEFTE , DIE IM AUFTRAG DER BUNDESREPUBLIK DEUTSCHLAND AN HOEHEREN SCHULEN DEUTSCH UNTERRICHTEN.

514

DAS AUSWAERTIGE AMT HAT MIT ZEHN AFRIKANISCHEN STAATEN ABKOMMEN GESCHLOSSEN , DIE DEN DEUTSCHUNTERRICHT IN DIESEN LAENDERN VOM SENEGAL BIS NACH MADAGASKAR SICHERN.

515
IN DEN MEISTEN DIESER STAATEN IST DEUTSCH IM RAHMEN DES BILDUNGSPLANS NACH ENGLISCH DIE ZWEITE FREMDSPRACHE

516
AN DER ELFENBEINKUESTE , IN TOGO UND IM SENEGAL HAT DEUTSCH LATEIN ALS SPRACHE DER LOGIKSCHULUNG VERDRAENGT , AN EINIGEN VERSUCHSGYMNASIEN WURDE ES GAR ERSTE FREMDSPRACHE.

517
DIESE AUFWAERTSENTWICKLUNG , DIE NIEMAND VORAUSGESEHEN HATTE , BERUHT AUF ZWEI FAKTEN : DEUTSCHLAND IST IM BEWUSSTSEIN DER AFRIKANER KEINE EHEMALIGE KOLONIALMACHT , DA ES SEINE KOLONIEN 1918 VERLOR , UND DER NIMBUS DER DEUTSCHEN TUECHTIGKEIT UND WIRTSCHAFTSKRAFT , AN EINIGEN IMPOSANTEN ENTWICKLUNGSPROJEKTEN VERANSCHAULICHT , IST UNGEBROCHEN.

518
AUCH HABEN DIE STAATSBESUCHE DES FRUEHEREN BUNDESPRAESIDENTEN HEINRICH LUEBKE IN EINIGEN ZENTRALAFRIKANISCHEN STAATEN ZU EINEM GROESSEREN SCHUELERANDRANG IM DEUTSCHUNTERRICHT GEFUEHRT.

519
ALS WAHLPFLICHFACH STEHT DEUTSCH IN DER REGEL MIT SPANISCH , RUSSISCH - UND IN MANCHEN STAATEN MIT MOHAMMEDANISCHER MEHRHEIT AUCH MIT ARABISCH - IM KONKURRENZKAMPF.

590
DIE SCHUELER , DIE EINE DIESER SPRACHEN WAEHLEN MUESSEN , ENTSCHEIDEN SICH IN IHRER UEBERWIEGENDEN MEHRHEIT FUER DEUTSCH , DAS HEISST , DIE ZAHL DER DEUTSCHSCHUELER IST MINDESTENS SO GROSS WIE DIE DER SCHUELER , DIE EINE DER DREI ANDEREN SPRACHEN GEWAEHLT HABEN , ZUSAMMEN.

521
FUER DIE DEUTSCHLEHRER IM FRANKOPHONEN AFRIKA BEDEUTET DAS GEGENUEBER IHREN RUSSISCHEN , FRANCOSPANISCHEN UND ARABISCHEN KOLLEGEN EINE WEITAUS GROESSERE BELASTUNG . DA SIE NICHT SELTEN KLASSEN BIS ZU FUENFZIG SCHUELERN UNTERRICHTEN.

522
BITTE LASSEN SIE SICH NICHT STOEREN , WENN DIESER SPIEGEL EINMAL KLIRRT.

523

ER IST SO ANGEBRACHT , DASS ER VOR DEN AUSWIRKUNGEN STARKER SCHALLWELLEN BEI DEN STARTS GROSSER RAKETEN GESCHUETZT IST.

<S. 190> TEST: ZEITUNGSARTIKEL SEITE 3

524
EINE KLEINE MIT SCHREIBMASCHINE GESCHRIEBENE NOTIZ FUER DIE GAESTE EINES MOTELS VON COCOA BEACH , ETWA 25 KILOMETER LUFTLINIE VON DER STARTRAMPE DER SATURN-5-TRAEGERRAKETE ENTFERNT.

525
WESENTLICH NAEHER AN DER STARTRAMPE , AUF DER FUENF KILOMETER ENTFERNTEN TRIBUENE IM KENNEDY-RAUMFUGZENTRUM , WIRD BEGREIFLICH , WARUM NOCH WEITAB VOM SCHUSS DIE SCHEIBEN KLIRREN KOENNEN.

526
DER BERSTENDE DONNER DER STARTENDEN SATURN 5 ERDRUECKT JEDEN ANDEREN LAUT.

527
IN EINER MINUTE IST ES JEDOCH VORBEI.

528
DIESE WELLE , DIE VON DER RAMPE UEBER DAS WASSER UND DIE FLACHEN LANDSTRICHE HERANROLLT , TRIFFT AUCH AUF HUNDERTTAUSENDE VON ZAUNGAESTEN , DIE AUS ALLEN RICHTUNGEN DES LANDES HERBEIKAMEN.

529
IN ZELTEN , LUXUSLIMOUSINEN UND ABENTEUERLICHEN VEHIKELN HATTEN SIE ES SICH SCHON AM SONNTAGABEND AN DEN STRASSENBOESCHUNGEN UND UFERN DES BANANA-RIVER UND DES INDIANRIVER , DIE AMERIKAS MONDFLUGHAFEN EINGRENZEN , BEQUEM GEMACHT.

530
TELEKAMERAS UND TELESKOPE WAREN AUF DAS FERNE TURMARTIGE GEBILDE APOLLO 15 GERICHTET.

531
BEI KLARER MORGENSICHT EINES TROPISCH-HEISSEN SOMMERTAGES KAMEN SIE WOHL ALLE AUF IHRE KOSTEN , DIE SCHLACHTENBUMMLER , DIE EHRENGAESTE , DIE OFFIZIELLEN UND DIE BEOBACHTENDEN JOURNALISTEN , DIE MITUNTER METERBREITE NACHRICHTENLABORS VOR SICH AUFGEBAUT HATTEN.

532
AUF EINER GLEISSEND HELLEN FLAMME VERSCHWAND APOLLO 15 AM HIMMEL.

533

ERST ALS DER DONNER LEISER WURDE , VEREBBTE DER JUBEL DER TAUSENDEN AM KAP.

534
DAS TAEGLICHE GEWITTER DIESER ZONE LIESS AN DIESEM TAG WENIGSTENS VORLAEUFIG AUF SICH WARTEN.

535
DAS UEBERRASCHENDE ERGEBNIS UNSERES TESTS GLEICH VORWEG: SAFTZENTRIFUGE UND SAFTAUTOMAT UNTERSCHEIDEN SICH IM WESENTLICHEN NUR NOCH DADURCH , DASS IM AUTOMATEN GROESSERE MENGEN OBST UND GEMUESE OHNE PAUSE HINTEREINANDER ENTSAFTET WERDEN KOENNEN.

536
BEIDE GERAETETYPEN LIEFERN GLEICH GUTEN SAFT IN GLEICHEN MENGEN.

<S. 191> TEST; ZEITUNGSARTIKEL SEITE 4

537
DASS EIN AUTOMAT PAUSENLOS HINTEREINANDER ENTSAFTEN KANN , LIEGT AN EINEM KLEINEN , ABER WESENTLICHEN UNTERSCHIED , DER ZWISCHEN IHM UND DER ZENTRIFUGE BESTEHT.

538
BEI BEIDEN GERAETEN WERDEN ZUM BEISPIEL ENTSPRECHEND ZERTEILTE AEPFEL IN DEN EINFUELLSTUTZEN GEGEBEN UND DURCH EINE SICH SCHNELL DREHENDE RASPELSCHEIBE ZERKLEINERT.

519
IN DER ZENTRIFUGE WIRD DER TRESTER - DAS SIND DIE RESTE VON SCHALEN UND FRUCHTFLEISCH - IN EINEM SIEBKORB GESAMMELT.

540
DIE ZENTRIFUGALKRAFT PRESST DEN SAFT DURCH DIE OEFFNUNGEN DES SIEBKORBES.

541
JE NACH SORTE UND GERAET MUSS BEI AEPFELN NACH ETWA 500 BIS 1000 GRAMM DER KORB GELEERT UND GESAEUBERT WERDEN.

542
BEI AUTOMATEN DAGEGEN BLEIBT DER TRESTER NICHT IM SIEBKORB , SONDERN WIRD IN EINEN AUFFANGBEHAELTER GESCHLEUDERT.

543
DIESER FASST DIE RUECKSTAENDE VON ETWA FUENF KILO AEPFELN.

544
EIN AUTOMAT KOMMT ALSO HAUPTSAECHLICH DANN FUER SIE IN FRAGE , WENN VIELE GROSSE SAFTTRINKER IN DER FAMILIE SIND ODER

WENN SAFT AUF FLASCHEN GEZOGEN UND FUER VITAMINAERMERE JAHRESZEITEN AUFGEHOBEN WERDEN SOLL.

545
REICHE WESTDEUTSCHE , DIE IN STEUEROASEN WIE DIE SCHWEIZ , LIECHTENSTEIN ODER DIE BAHAMAS UEBERWECHSELN , MUESSEN KUENFTIG ZEHN JAHRE LANG FUER IHRE BISHER KAUM BESTEUERTEN EINKUENFTE UND VERMOEGEN IN DER BUNDESREPUBLIK DIE VOLLE STEUERLAST TRAGEN.

546
NACH DEM GESETZENTWURF ZUR BEKAEMPFUNG DER STEUERFLUCHT , DEN DAS BONNER KABINETT AM VERGANGENEN MITTWOCH VERABSCHIEDET HAT , GILT DIE ZEHNJAHRESFRIST AUCH FUER DIEJENIGEN , DIE SCHON FRUEHER IHREN WOHNSITZ VERLEGT HABEN.

547
DAMIT KOENNEN DIE FINANZAEMTER ZUMINDEST FUER EINIGE JAHRE AUCH JENE RUND 500 MILLIONAERE ERFASSEN , DIE SICH VOR ALLEM SEIT 1966-1967 IN STEUEROASEN ABGESETZT HABEN

548
BESTEUERT WERDEN NACH DEM REGIERUNGSENTWURF VERMOEGEN VON UEBER 300000 MARK UND EINKUENFTE VON MEHR ALS 120000 MARK IM JAHR.

549
FUER DIE HOEHE DES STEUERSATZES IST ALLERDINGS DAS SORENANNTE WELTEINKOMMEN DES NUNMEHR STEUERPFLICHTIGEN STEUERFLEUCHTLINGS MASSGEBEND.

<S. 192> <ZEITUNGSTEXTE> SEITE 5

550
WENN DAS STEUEROASENGESETZ - RUECKWIRKEND - AM ERSTEN ERSTEN 1971 IN KRAFT TRETEN SOLL , MUSS ES NOCH IN DIESEM JAHR VON BUNDESTAG UND BUNDESRAT VERABSCHIEDET WERDEN.

551
SICHERHEIT AUF DER STRASSE WIRD ZUR EXAKTEN WISSENSCHAFT.

552
LICHTINGENIEURE VON SIEMENS HABEN DIE ENTSCHEIDENDE FORMEL GEFUNDEN , DAS QQ-VERFAHREN ERMOEGLICHT DIE LICHTTECHNISCME KENNZEICHNUNG JEDER BELIEBIGEN STRASSENDECKE UND DAMIT DIE VORAUSBERECHNUNG DER LEUCHTDICHTE UND LEUCHTDICHTEGLEICHMAESSIGKEIT.

553
DIESE MESSDATEN BILDEN BEREITS HEUTE DEN INTERNATIONAL ANERKANNTEN NORMMASSSTAB IN DER STRASSENBELEUCHTUNG.

554

COMPUTER KOENNEN JETZT FUER KOMPLIZIERTE LEUCHTDICHTEBERECHNUNGEN EINGESETZT WERDEN.

595
DIE ERGEBNISSE SIND IN MINUTEN VERFUEGBAR - PRAKTISCH AUF KNOPFDRUCK.

596
SIEMENS-LICHTINGENIEURE KOENNEN AUF EINE UEBER 100-JAEHRIGE TAETIGKEIT IN BEREICH DER ZWECKLEUCHTENFORSCHUNG AUFBAUEN
-
EINE GUTE VORAUSSETZUNG , ZUKUNFTSWEISENDE TECHNIKEN FUER FUNKTIONSGERECHTES LICHT ZU ENTWICKELN.

*

<Umschlagtext Rückseite>

Ein Lexikon in Buchform ist unbeweglich, es trägt den Veränderungen des Wortschatzes nur ungenügend Rechnung (etwa durch Modifikation in Neuauflagen). Der Computer bietet die Möglichkeit, von einem statischen Lexikon überzugehen zu einem dynamischen System, das sich (anhand von Texten) selbständig den Gegebenheiten der Sprache anpaßt. Fragen der Lexikon-Optimierung (Basislexikon), der automatischen kontextsensitiven Klassifikation nichtinventarisierter Wörter und des Aufbaus eines maschinellen Lernverfahrens zur kurzfristigen, textbezogenen Lexikonergänzung (Kurzzeitlexikon) stehen im Mittelpunkt aller Untersuchung.

Der Autor ist Mitarbeiter des Institutes für linguistische Datenverarbeitung, Saarbrücken.